

HỌC VIỆN NGÂN HÀNG
KHOA HỆ THỐNG THÔNG TIN QUẢN LÝ



BÁO CÁO BÀI TẬP LỚN

Học phần: Hệ hỗ trợ ra quyết định và kinh doanh thông minh

Đề tài:

**XÂY DỰNG KHO ĐỮA LIỆU NÂNG CAO HIỆU QUẢ
QUẢN LÝ VÀ BÁO CÁO PHÂN TÍCH THỐNG KÊ CHO
CÔNG TY CỔ PHẦN CÔNG NGHỆ GIÁO DỤC EDTE**

Giảng viên hướng dẫn : ThS. Ngô Thùy Linh

Mã lớp học phần : 221IS31A01

Nhóm sinh viên thực hiện : Nhóm 4

Nguyễn Thị Minh Nguyệt – 22A4040031

Bùi Thị Bích Hằng – 22A4040026

Trần Thị Hải Yến – 22A4040002

Nguyễn Thị Linh – 22A4040107

Phạm Văn Đông – 23A4040027

Hà Nội – 2022

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

THÔNG TIN CHUNG

TÊN ĐỀ TÀI: Xây dựng kho dữ liệu nâng cao hiệu quả quản lý và báo cáo phân tích thống kê cho công ty cổ phần công nghệ giáo dục EDTE

Tổ chức nghiên cứu: CÔNG TY CỔ PHẦN CÔNG NGHỆ GIÁO DỤC EDTE

Danh sách thành viên nhóm 4:

S T T	Thành viên	Mã sinh viên	Nội dung công việc	Đóng góp (%)
1	Nguyễn Thị Minh Nguyệt	22A4040031	Tổng quan về Data Lake Chương 6, 8	20%
2	Bùi Thị Bích Hằng	22A4040026	Chương 5, 8	20%
3	Trần Thị Hải Yến	22A4040002	Chương 7 Tổng hợp, chỉnh sửa word	20%
4	Nguyễn Thị Linh	22A4040107	Hạn chế của kho dữ liệu, Kết luận	20%
5	Phạm Văn Đông	23A4040027	Lời mở đầu, Chương 7	20%

Giảng viên hướng dẫn: Ths. Ngô Thùy Linh.

LỜI CẢM ƠN

Chúng em xin gửi lời cảm ơn chân thành tới cô Ngô Thùy Linh - Giảng viên Khoa Hệ thống thông tin quản lý, Học viện Ngân hàng. Trong quá trình học tập và thực hiện bài báo cáo “Xây dựng kho dữ liệu nâng cao hiệu quả quản lý và báo cáo phân tích thống kê cho công ty cổ phần công nghệ giáo dục EDTE”, cô đã luôn tạo điều kiện, giúp đỡ để chúng em hoàn thành được bài báo cáo này.

Em xin chân thành gửi lời cảm ơn đến ban lãnh đạo và các anh/chị trong Công ty cổ phần công nghệ giáo dục EDTE đã tạo điều kiện thuận lợi trong suốt thời gian em làm việc tại công ty và cung cấp nguồn dữ liệu giúp bài làm của nhóm có tính thực tế cao. Việc được tiếp xúc thực tế, hướng dẫn tường tận đã giúp em có thêm hiểu biết và được cọ xát với môi trường làm việc thực tế.

Nhóm đã cố gắng hoàn thiện bài báo cáo với tất cả sự nỗ lực và cố gắng của cả nhóm. Tuy nhiên, do còn thiếu nhiều kinh nghiệm, chắc chắn bài báo cáo sẽ không tránh khỏi thiếu sót. Vì vậy, chúng em rất mong nhận được sự quan tâm, những ý kiến đóng góp của cô để bài báo cáo của chúng em có thể hoàn thiện hơn.

Chúng em xin chân thành cảm ơn!.

MỤC LỤC

NHẬN XÉT CỦA GIÁNG VIÊN HƯỚNG DẪN	ii
THÔNG TIN CHUNG	iii
LỜI CẢM ƠN.....	iv
MỤC LỤC	v
DANH MỤC HÌNH ẢNH, BẢNG BIỂU	x
LỜI MỞ ĐẦU	1
CHƯƠNG 1. CƠ SỞ LÝ THUYẾT.....	2
1.1. Tổng quan về kho dữ liệu (Data warehouse)	2
1.1.1. Khái niệm kho dữ liệu	2
1.1.2. Sự khác nhau giữa CSDL và KDL	2
1.1.3. Phân loại kho dữ liệu	3
1.1.4. Các cách tiếp cận xây dựng kho dữ liệu.....	4
1.1.5. Hạn chế của kho dữ liệu	8
1.2. Tích hợp dữ liệu.....	8
1.2.1. Lý do phải tích hợp dữ liệu từ nhiều nguồn	8
1.2.2. Các vấn đề liên quan đến tích hợp dữ liệu	10
1.3. Quản trị và quản lý dữ liệu	12
1.3.1. Phân biệt quản trị và quản lý dữ liệu	12
1.3.2. Lý do phải quản lý dữ liệu chủ.....	14
1.4. Tổng quan về hồ dữ liệu (Data Lake)	15
1.4.1. Khái niệm hồ dữ liệu	15
1.4.2. Kiến trúc Data Lake	16
1.4.3. Ưu điểm và hạn chế của Data Lake trong quá trình sử dụng	17
1.4.3.1. Ưu điểm.....	17
1.4.3.2. Hạn chế.....	18

1.4.4. So sánh Data Warehouse và Data Lake	19
CHƯƠNG 2. TỔNG QUAN VỀ CÔNG TY	21
2.1. Giới thiệu về công ty.....	21
2.2. Hiện trạng	21
2.3. Quy trình phân tích	22
2.3.1. Quy trình quản lý học tập	24
2.3.2. Quy trình quản lý bán hàng	26
CHƯƠNG 3. GIỚI THIỆU VÀ XÂY DỰNG CÁC HỆ THỐNG NGUỒN DỮ LIỆU	
27	
3.1. Hệ thống nguồn 1 (SQL Server - Phòng đào tạo).....	27
3.1.1. Thiết kế cơ sở dữ liệu mức khái niệm	27
3.1.1.1. Xác định thực thể	27
1.1.1.1. Xác định mối quan hệ thực thể.....	27
3.1.1.2. Thiết kế cơ sở dữ liệu mức logic	27
3.1.2.1. Chuyển hóa quan hệ thành các quan hệ	27
3.1.2.2. Chuẩn hóa thực thể.....	28
3.1.1.3. Thiết kế cơ sở dữ liệu mức vật lý	28
3.1.4. Mô hình dữ liệu quan hệ đã chuẩn hóa	32
3.1.5. Trigger tự động tính số buổi còn lại	32
3.1.6. Dữ liệu nguồn 1	33
3.2. Hệ thống nguồn 2 (Excel File - Phòng Sale)	33
3.3. Hệ thống nguồn 3 (MySQL Source).....	34
CHƯƠNG 4. THIẾT KẾ VÀ MÔ TẢ CÁC BẢNG DIM, FACT	36
4.1. Nghiệp vụ chính gắn với các bảng Dim, Fact trong Data Mart.....	36
4.2. Lựa chọn kiểu thiết kế phù hợp cho bảng Fact.....	36
4.3. Mô tả rõ ý nghĩa các trường dữ liệu và nguồn của mỗi bảng Dim, Fact	36

4.3.1.	Xây dựng bảng cắt lớp thời gian Dim_Date	36
4.3.2.	Xây dựng bảng cắt lớp Học viên (Dim_HocVien).....	47
4.3.3.	Xây dựng bảng cắt lớp nhân viên Sale (Dim_NVSale)	55
4.3.4.	Xây dựng bảng cắt lớp giáo viên (Dim_GiaoVien)	56
4.3.5.	Xây dựng bảng cắt lớp khóa học (Dim_KhoaHoc).....	58
4.3.6.	Xây dựng bảng cắt lớp Đăng Ký (Dim_DangKy)	60
4.3.7.	Xây dựng bảng cắt lớp Thanh Toán (Dim_ThanhToan).....	61
4.3.8.	Xây dựng bảng cắt lớp Học (Dim_Hoc)	63
4.3.9.	Xây dựng bảng Fact_QLHoc	64
4.3.10.	Xây dựng bảng Fact_Sales	66
4.3.11.	Xây dựng sơ đồ chòm sao	68
4.4.	Cập nhập dữ liệu vào kho dữ liệu	68
4.4.1.	Xây dựng flow	68
4.4.2.	Data flow bảng Dim_NVsales, Dim_GiaoVien, Dim_KhoaHoc, Dim_DangKy, Dim_ThanhToan, Dim_Hoc	69
4.4.3.	Data flow bảng Fact_Sales	69
4.4.4.	Data flow bảng QLHoc	70
	CHƯƠNG 5. CẬP NHẬT DỮ LIỆU BẢNG DIM	72
5.1.	Ba kiểu cập nhật dữ liệu trên bảng Dimension	72
5.1.1.	Kiểu SCD1 (Type 1 Slowly Changing Dimension)	72
5.1.2.	Kiểu SCD2 (Type 2 Slowly Changing Dimension)	73
5.1.3.	Kiểu SCD3 (Type 3 Slowly Changing Dimension)	74
5.2.	Thực hành cập nhật dữ liệu trên bảng Dimension	74
	CHƯƠNG 6. XÂY DỰNG VÀ TRIỂN KHAI DATA LAKE TRÊN AMAZON WEB SERVICE (AWS)	81
6.1.	Tìm hiểu giải pháp xây dựng Data Lake thu thập dữ liệu phi cấu trúc cho doanh nghiệp EDTE	81

6.2. Đề xuất doanh nghiệp EDTE nên nâng cấp hệ thống hiện tại lên Cloud và ứng dụng giải pháp Data Lake trên AWS.....	82
6.2.1. Lý do doanh nghiệp cần lên kế hoạch nâng cấp hệ thống hiện tại lên Cloud và ứng dụng Data Lake	82
6.2.2. Lợi ích doanh nghiệp sẽ nhận được nếu nâng cấp hệ thống hiện tại lên Cloud và ứng dụng Data Lake	83
6.3. Xây dựng và triển khai Data Lake trên Amazon Web Services	84
6.3.1. Các dịch vụ được sử dụng trên AWS để xây dựng Data Lake.....	84
6.3.2. Thực hành xây dựng Data Lake trên AWS	85
CHƯƠNG 7. XÂY DỰNG CÁC BÁO CÁO PHÂN TÍCH THỐNG KÊ.....	118
7.1. Xác định BI User	118
7.2. Xây dựng các báo cáo phân tích thống kê (dashboard).....	119
7.2.1. Bảng câu hỏi	119
7.2.2. Lập báo cáo.....	120
7.2.2.1. Báo cáo tổng quan tình hình kinh doanh.....	121
7.2.2.2. Báo cáo quản lý đào tạo	122
7.2.2.3. Báo cáo doanh thu phòng Sales	123
7.2.2.4. Báo cáo doanh thu tại phòng kế toán	124
7.2.2.5. Báo cáo lượng khóa học đăng ký vào các đợt khuyến mại	125
CHƯƠNG 8. ỨNG DỤNG MACHINE LEARNING TRONG XÂY DỰNG MÔ HÌNH ĐUỢC ĐOÁN HỌC VIÊN TÁI TỤC.....	126
8.1. Tổng quan về Machine Learning (Học máy).....	126
8.1.1. Khái niệm Machine Learning.....	126
8.1.2. Phân loại	126
8.1.3. Quy trình xây dựng mô hình machine learning.....	126
8.1.4. Các ứng dụng của học máy	127

8.2. Giới thiệu về mô hình cây quyết định (Decision Tree) và thuật toán rừng ngẫu nhiên (Random forest)	127
8.2.1. Mô hình Decision Tree (Cây quyết định).....	127
8.2.2. Thuật toán Random Forest (Rừng ngẫu nhiên)	129
8.3. Tính cấp thiết trong xây dựng mô hình dự báo học viên tái tục	131
8.4. Xây dựng mô hình dự báo tái tục	132
8.4.1. Thu thập dữ liệu.....	132
8.4.2. Xử lý dữ liệu.....	133
8.4.3. Xây dựng mô hình dự đoán tái tục với hai thuật toán Rừng ngẫu nhiên và Cây quyết định	134
8.4.3.1. Mô hình Decision Tree.....	136
8.4.3.2. Thuật toán Random Forest	137
8.5. Kiểm thử mô hình dự đoán với Decision True và Random Forest.....	139
8.6. Kiểm tra mô hình và đánh giá độ chính xác của mô hình	140
8.6.1. Độ chính xác của mô hình	140
8.6.2. Ma trận nhầm lẫn.....	141
KẾT LUẬN	144
TÀI LIỆU THAM KHẢO	146
PHỤ LỤC	148
Phụ lục 1: Nguồn dữ liệu mà nhóm tổng hợp được từ công ty	148
Phụ lục 2: Hướng dẫn các thao tác cài đặt và kết nối với dịch vụ tích hợp máy chủ SQL (SQL Server Integration Services) khi sử dụng MySQL.	156
Phụ lục 3: Khai phá Dữ liệu	159

DANH MỤC HÌNH ẢNH

Hình 1.1 Các vấn đề trong tích hợp dữ liệu.....	10
Hình 1.2 Hồ dữ liệu (Nguồn: Amazon Web Services).....	16
Hình 1.3 Kiến trúc Data Lake (Nguồn: guru99.com).....	16
Hình 1.4 Sự khác nhau giữa Data Warehouse và Data Lake (Nguồn: grazitti.com) ...	19
Hình 2.1 Sơ đồ cơ cấu tổ chức công ty cổ phần giáo dục EDTE	21
Hình 2.2 Quy trình làm việc giữa các phòng ban tại công ty Edura (Nguồn: Công ty Edura)	23
Hình 2.3 Chức năng của mỗi phòng (Nguồn: Công ty Edura).....	24
Hình 2.4 Quy trình quản lý học viên (Nguồn: Công ty Edura)	24
Hình 2.6 Quy trình vận hành lớp học (Nguồn: Công ty Edura)	25
Hình 2.7 Quy trình chăm sóc học viên (Nguồn: Công ty Edura)	25
Hình 3.1 Mô hình dữ liệu quan hệ đã chuẩn hóa.....	32
Hình 3.2 Nguồn dữ liệu 1	33
Hình 3.3 Dữ liệu trên Excel.....	34
Hình 3.4 Hệ thống nguồn thứ 3 (MySQL Source)	35
Hình 4.1 Bảng Dim_Date	37
Hình 4.2 Kết quả sau khi thực thi câu lệnh Insert bảng Dim_Date	47
Hình 4.3 Bảng Dim_HocVien	48
Hình 4.4 Nguồn 1 (SQL Server source)	48
Hình 4.5 Nguồn 2 (File Excel)	49
Hình 4.6 Nguồn 3 (MySQL source)	50
Hình 4.7 Thực hiện đổ dữ liệu từ các nguồn tới bảng Dim_NVSale	54
Hình 4.8 Kết quả thu được tại bảng Dim_HocVien	54
Hình 4.9 Bảng Dim_NVSale	55
Hình 4.10 Thực hiện đổ dữ liệu tới bảng Dim_NVSale.....	55
Hình 4.11 Kết quả thu được tại bảng Dim_NVSale.....	56
Hình 4.12 Bảng Dim_GiaoVien	56
Hình 4.13 Thực hiện đổ dữ liệu từ nguồn 1 tới bảng Dim_GiaoVien	57

Hình 4.14 Kết quả thu được tại bảng Dim_GiaoVien.....	57
Hình 4.15 Bảng Dim_KhoaHoc	58
Hình 4.16 Thực hiện đồ dữ liệu cho bảng Dim_KhoaHoc.....	59
Hình 4.17 Kết quả thu được tại bảng Dim_KhoaHoc	59
Hình 4.18 Bảng Dim_DangKy	60
Hình 4.19 Thực hiện đồ dữ liệu tới bảng Dim_DangKy.....	60
Hình 4.20 Kết quả thu được tại bảng Dim_DangKy.....	61
Hình 4.21 Bảng Dim_ThanhToan	62
Hình 4.22 Thực hiện đồ dữ liệu tới bảng Dim_ThanhToan.....	62
Hình 4.23 Kết quả thu được tại bảng Dim_ThanhToan.....	63
Hình 4.24 Thực hiện đồ dữ liệu tới bảng Dim_Hoc.....	64
Hình 4.25 Kết quả thu được tại bảng Dim_Hoc	64
Hình 4.26 Bảng Fact_QLHoc.....	65
Hình 4.27 Bảng Fact_Sales	67
Hình 4.28 Sơ đồ chòm sao.....	68
Hình 4.29 Data Flow	69
Hình 4.30 Data flow bảng Dim_NVsales, Dim_GiaoVien, Dim_KhoaHoc, Dim_DangKy, Dim_ThanhToan, Dim_Hoc	69
Hình 4.31 Data flow bảng Fact_Sales	70
Hình 4.32 Kết quả thu được tại bảng Fact_Sales	70
Hình 4.33 Data flow bảng Fact_QLHoc.....	71
Hình 4.34 Kết quả thu được tại bảng Fact_QLHoc.....	71
Hình 5.1 Minh họa SCD kiểu 1 (Nguồn: Understanding Slowly Changing Dimensions, Oracle)	72
Hình 5.2 Minh họa SCD kiểu 2 (Nguồn: Understanding Slowly Changing Dimensions, Oracle)	73
Hình 6.1 Sơ đồ minh họa kiến trúc hồ dữ liệu sử dụng trên AWS.....	84
Hình 7.1 Báo cáo tổng quan tình hình kinh doanh	121
Hình 7.2 Báo cáo quản lý đào tạo.....	122
Hình 7.3 Báo cáo doanh thu phòng Sales.....	123

Hình 7.4 Báo cáo doanh thu tại phòng kế toán	124
Hình 7.5 Báo cáo lượng khóa học đăng ký vào các đợt khuyến mại	125
Hình 8.1 Cấu trúc của 1 cây quyết định (Nguồn: Decision Tree Algorithms-Machine Learning, Rupika Nimbalkar).....	128
Hình 8.2 Các bước hoạt động của thuật toán Random Forest (Nguồn: Random Forests Classifiers, Shivam Sharma)	130
Hình 8.3 Chi tiết quy trình chăm sóc học viên tái tục (Nguồn: Công ty Edura).....	131
Hình 8.4 Dữ liệu thu thập được	132
Hình 8.5 Dữ liệu sau khi đã làm sạch.....	133
Hình 8.6 Hình ảnh cây quyết định sau khi chạy mô hình Decision Tree.....	137
Hình 8.7 Hình ảnh cây quyết định sau khi chạy thuật toán Random Forest	138
Hình 8.8 Kết quả chạy mô hình dự đoán với mô hình Decision Tree.....	139
Hình 8.9 Kết quả chạy mô hình dự đoán với mô hình Random Forest.....	140
Hình 8.10 Ma trận nhầm lẫn mô hình Decision Tree	141
Hình 8.11 Kết quả độ đo đánh giá của bài toán với mô hình Decision Tree	141
Hình 8.12 Ma trận nhầm lẫn thuật toán Random Forest	142
Hình 8.13 Kết quả độ đo đánh giá của bài toán với mô hình Random Forest	142
Hình 8.14 So sánh ma trận nhầm lẫn giữa Decision Tree và Random Forest.....	143

DANH MỤC BẢNG BIỂU

Bảng 1.1 Sự khác nhau giữa Cơ sở dữ liệu và Kho dữ liệu	3
Bảng 1.2 Ba loại kho dữ liệu chính	4
Bảng 1.3 Sự khác nhau giữa cách tiếp cận Top-down và Bottom-up	7
Bảng 1.4 So sánh Quản trị dữ liệu và Quản lý dữ liệu	13
Bảng 1.5 Mô tả sự khác nhau giữa Data Lake và Data warehouse (Nguồn: guru99.com)	20
Bảng 3.1 Cơ sở dữ liệu mức vật lý bảng Học Viên.....	28
Bảng 3.2 Cơ sở dữ liệu mức vật lý bảng Giáo Viên.....	29
Bảng 3.3 Cơ sở dữ liệu mức vật lý bảng Khóa Học	29
Bảng 3.4 Cơ sở dữ liệu mức vật lý bảng Nhân Viên Sale	30

Bảng 3.5 Cơ sở dữ liệu mức vật lý bảng Đăng ký	30
Bảng 3.6 Cơ sở dữ liệu mức vật lý bảng Học	31
Bảng 3.7 Cơ sở dữ liệu mức vật lý bảng Thanh Toán.....	31
Bảng 4.1 Các trường dữ liệu của bảng Fact_QLHoc	66
Bảng 4.2 Các trường dữ liệu của bảng Fact_Sales.....	68
Bảng 7.1 BI User của công ty.....	119
Bảng 7.2 Bảng câu hỏi.....	120
Bảng 8.1 Các tiêu chí làm sạch và xử lý dữ liệu	133
Bảng 8.2 Mô tả dữ liệu đầu vào cho quá trình kiểm thử mô hình dự đoán.....	139
Bảng 8.3 So sánh độ chính xác giữa Decision Tree và Random Forest.....	140

LỜI MỞ ĐẦU

Đối với một tổ chức, doanh nghiệp trong suốt thời gian hoạt động của mình đều phải đối mặt trước rất nhiều cơ hội, thách thức từ các yếu tố bên trong đến các yếu tố bên ngoài của doanh nghiệp. Đứng trước những cơ hội, thách thức đó sẽ là những quyết định vô cùng khó khăn của các nhà quản trị. Những quyết định đòi hỏi các nhà quản trị phải đưa ra một cách nhanh chóng, chính xác để góp phần giúp doanh nghiệp có thể tồn tại và thúc đẩy sự phát triển lớn mạnh của doanh nghiệp. Những nhà quản trị của tổ chức, doanh nghiệp mang trong mình sứ mệnh và trọng trách rất lớn với các quyết định của mình đưa ra bởi có những quyết định rất quan trọng, nguy hiểm có thể ảnh hưởng đến lợi ích trước mắt của doanh nghiệp hoặc lớn hơn là sự tồn vong của tổ chức, doanh nghiệp. Từ những tính chất quan trọng của việc ra quyết định của các nhà quản trị thì những nhà quản trị đòi hỏi cần một hệ thống có thể giúp họ ra những quyết định nhanh chóng, chính xác với một mức độ rủi ro thấp nhất. Từ đó nhiều hệ hỗ trợ ra quyết định đã được ra đời đi kèm với những sứ mệnh cao cả đồng hành không thể thiếu của những nhà quản trị. Đó là một thành phần quan trọng trong hệ thống quản trị doanh nghiệp thông minh – Business Intelligence (BI).

Hệ hỗ trợ ra đời nó đã tiết kiệm được khá nhiều thời gian bởi tốc độ xử lý của nó rất nhanh đi kèm với một lượng chi phí bỏ ra cũng là rất thấp để cho ra những báo cáo. Những báo cáo, thống kê cần các nhà quản trị phải bỏ thêm thời gian và trí óc để nghiên cứu đánh giá những số liệu trên những báo cáo đó để đưa ra những quyết định chính xác, kịp thời. Nó thích hợp cho các quyết định kinh doanh dựa trên số liệu thu thập được trong quá trình hoạt động của tổ chức, doanh nghiệp. Dữ liệu được cập nhập liên tục và những nhà quản trị có thể đưa ra những quyết định tức thời. Để hoạt động được hiệu quả hệ hỗ trợ cần kết hợp với một Data Warehouse (kho dữ liệu) - nơi chứa những dữ liệu cần thiết, dữ liệu đã được làm sạch phục vụ cho quá trình tạo các báo cáo, thống kê, bảng biểu. Hệ hỗ trợ và Data Warehouse có mối quan hệ mật thiết tạo nên một hệ hỗ trợ ra quyết định mạnh mẽ, hiệu quả và đơn giản. Nghiên cứu này sẽ tập trung nghiên cứu và đưa ra quyết định kinh doanh của công ty cổ phần công nghệ giáo dục EDTE.

CHƯƠNG 1. CƠ SỞ LÝ THUYẾT

1.1. Tổng quan về kho dữ liệu (Data warehouse)

1.1.1. Khái niệm kho dữ liệu

- Kho dữ liệu (Data warehouse) là nơi lưu trữ dữ liệu bằng thiết bị điện tử của một tổ chức, doanh nghiệp, nhằm hỗ trợ việc phân tích dữ liệu và lập báo cáo.
- Cơ sở dữ liệu (Database) là tập hợp các dữ liệu có cấu trúc, có tổ chức và mối liên quan với nhau, thường được lưu trữ và truy cập điện tử từ hệ thống máy tính, được nhiều người sử dụng và được tổ chức theo một mô hình.

1.1.2. Sự khác nhau giữa Cơ sở dữ liệu và Kho dữ liệu

Tham số	Cơ sở dữ liệu (Database)	Kho dữ liệu (Data Warehouse)
Mục đích	Để ghi và truy vấn dữ liệu	Để xử lý và phân tích dữ liệu
Chức năng	Hỗ trợ các hoạt động cơ bản hàng ngày	Hỗ trợ quyết định mang tính chiến lược, đưa ra các dự án
Phương pháp xử lý	Cơ sở dữ liệu sử dụng Xử lý giao dịch trực tuyến (OLTP)	Kho dữ liệu sử dụng Xử lý phân tích trực tuyến (OLAP)
Các bảng và phép nối	Có độ phức tạp cao vì chúng được chuẩn hóa (cho RDMS) để giảm dữ liệu thừa, tối ưu hóa dung lượng lưu trữ	Bảng và phép nối rất dễ dàng trong kho dữ liệu vì chúng không được chuẩn hóa
Tính chất dữ liệu	Chi tiết, được cập nhật thường xuyên	Có tính lịch sử và thống kê, được thêm mới chứ không cập nhật
Loại dữ liệu	Dữ liệu được lưu trữ trong Cơ sở dữ liệu được cập nhật	Dữ liệu hiện tại và lịch sử được lưu trữ trong Kho dữ liệu. Có thể không được cập nhật
Lưu trữ dữ liệu	Phương pháp tiếp cận quan hệ phẳng, nhiều dữ liệu khác	Phương pháp tiếp cận đa chiều và chuẩn hóa, nhiều nguồn dữ liệu

	nhau được tích hợp vào một nguồn	khác nhau được tích hợp và định dạng lại
Tóm tắt dữ liệu	Dữ liệu chi tiết được lưu trữ trong cơ sở dữ liệu	Lưu trữ dữ liệu tóm tắt tối đa
Mô hình thiết kế	Kỹ thuật mô hình quan hệ thực thể được sử dụng để thiết kế	Kỹ thuật mô hình dữ liệu được sử dụng để thiết kế
Kỹ thuật	Capture dữ liệu	Phân tích dữ liệu
Sử dụng	Thường xuyên	Trong những trường hợp đặc biệt
Đơn vị công việc	Giao dịch đơn giản, ngắn	Các câu truy vấn phức tạp
Độ đo	Thông lượng giao dịch, có thể thực hiện nhiều giao dịch cùng một lúc	Thông lượng truy vấn và trả lời
Sự định hướng	Định hướng ứng dụng	Định hướng chủ đề
Mô hình sử dụng	Mô hình quan hệ – thực thể	Mô hình dữ liệu đa chiều
Loại truy vấn	Những truy vấn giao dịch đơn giản được sử dụng	Những truy vấn phức tạp được áp dụng cho mục đích phân tích
Hiệu suất truy vấn phân tích	Thấp	Cao

Bảng 1.1 Sự khác nhau giữa Cơ sở dữ liệu và Kho dữ liệu

1.1.3. Phân loại kho dữ liệu

Có ba loại kho dữ liệu chính: Kho dữ liệu doanh nghiệp (EDW), Kho dữ liệu hoạt động (ODS) và Kho dữ liệu cục bộ (Data mart) (Singh, 2021).

Kho dữ liệu doanh nghiệp (EDW)	Kho dữ liệu hoạt động (ODS)	Kho dữ liệu cục bộ (Data mart)
Kho dữ liệu doanh nghiệp (EDW) là một kho tập trung cung cấp các dịch vụ hỗ trợ ra quyết định trong toàn doanh nghiệp. EDW thường là một tập hợp các cơ sở dữ liệu cung cấp một cách tiếp cận thống nhất để tổ chức dữ liệu và phân loại dữ liệu theo chủ đề. Mặt khác, EDW được sử dụng để hỗ trợ các quyết định chiến thuật và chiến lược.	Kho dữ liệu hoạt động (ODS) là cơ sở dữ liệu trung tâm được sử dụng để báo cáo hoạt động như một nguồn dữ liệu cho kho dữ liệu doanh nghiệp (EDW). ODS là một yếu tố bổ sung cho EDW và được sử dụng để báo cáo hoạt động, kiểm soát và ra quyết định. ODS được làm mới theo thời gian thực, khiến nó thích hợp hơn cho các hoạt động thường ngày như lưu trữ hồ sơ nhân viên.	Data mart được coi là một tập hợp con của kho dữ liệu và thường hướng đến một nhóm hoặc ngành kinh doanh cụ thể, chẳng hạn như tài chính hoặc bán hàng. Data mart hướng chủ đề, lưu trữ từng loại dữ liệu cụ thể có sẵn cho một nhóm người dùng xác định, cung cấp thông tin chi tiết, quan trọng. Sự sẵn có của dữ liệu này giúp người dùng không lãng phí thời gian tìm kiếm trong toàn bộ kho dữ liệu.

Bảng 1.2 Ba loại kho dữ liệu chính

1.1.4. Các cách tiếp cận xây dựng kho dữ liệu

Có 2 cách tiếp cận để xây dựng kho dữ liệu: Cách tiếp cận từ trên xuống (Top-down) và cách tiếp cận từ dưới lên (Bottom-Up) được giải thích như dưới đây.

- Phương pháp tiếp cận từ trên xuống – Top-down

Cách tiếp cận này được Inmon định nghĩa: Data Warehouse như một kho lưu trữ trung tâm cho toàn bộ tổ chức và các Data mart được tạo từ nó sau khi Data Warehouse hoàn chỉnh đã được tạo.

- Phương pháp tiếp cận từ dưới lên – Bottom-Up

Cách tiếp cận từ dưới lên được Kinball tạo ra với cách thức: Data mart được tạo trước và cung cấp một cái nhìn hẹp để các doanh nghiệp có thể phân tích và Data Warehouse được tạo ra sau khi tạo xong Data mart.

- Sự khác nhau giữa cách tiếp cận:

Tham số	Cách thực hiện	Ưu điểm	Nhược điểm
Top-down	<ul style="list-style-type: none"> - External Sources - Nguồn bên ngoài là nguồn từ đó dữ liệu được thu thập bất kể loại dữ liệu. Dữ liệu có thể có cấu trúc, bán cấu trúc và không cấu trúc. - Staging Area - Vùng giai đoạn: Vì dữ liệu được trích xuất từ các nguồn bên ngoài không tuân theo một định dạng cụ thể, vì vậy cần phải xác thực dữ liệu này để tải vào Data Warehouse. Với mục đích này, bạn nên sử dụng công cụ ETL. - E (Trích xuất): Dữ liệu được trích xuất từ Nguồn dữ liệu bên ngoài. 	<p>Cung cấp chế độ xem theo chiều nhất quán về các ô chứa dữ liệu.</p> <p>- Ngoài ra, mô hình này được coi là mô hình mạnh nhất cho những thay đổi trong kinh doanh. Đó là lý do hiện tại các tổ chức lớn thích làm theo cách tiếp cận này (Geeksforgeeks, 2021).</p> <p>- Dễ dàng tạo Data Mart từ Data Warehouse.</p>	<ul style="list-style-type: none"> - Chi phí bảo trì cao - Mất nhiều thời gian thiết kế.

	<p>- T (Transform): Dữ liệu được chuyển đổi sang định dạng chuẩn.</p> <p>- L (Load): Dữ liệu được tải vào Data Warehouse sau khi chuyển đổi nó thành định dạng chuẩn.</p> <p>- Data Warehouse - Kho dữ liệu: Sau khi làm sạch dữ liệu, nó được lưu trữ trong Data warehouse như một kho lưu trữ trung tâm. Nó thực sự lưu trữ dữ liệu meta và dữ liệu thực tế được lưu trữ trong các ngăn chứa dữ liệu. Lưu ý rằng Data Warehouse lưu trữ dữ liệu ở dạng tinh khiết nhất trong cách tiếp cận từ trên xuống này.</p> <p>- Data Mart: Data mart cũng là một phần của thành phần lưu trữ. Nó lưu trữ thông tin về một chức năng cụ thể của một tổ chức được xử lý bởi một cơ quan duy nhất. Có thể có bao nhiêu số lượng Data mart trong một tổ chức tùy</p>	
--	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--

	<p>thuộc vào các chức năng.</p> <p>Chúng ta cũng có thể nói rằng Data mart chứa tập hợp con của dữ liệu được lưu trữ trong Data warehouse.</p> <ul style="list-style-type: none"> - Data Mining - Khai phá dữ liệu: Thực hành phân tích dữ liệu lớn có trong data warehouse là khai thác dữ liệu. Nó được sử dụng để tìm các mẫu ẩn có trong cơ sở dữ liệu hoặc trong Data warehouse với sự trợ giúp của thuật toán khai thác dữ liệu. 		
Bottom-up	<ul style="list-style-type: none"> - Đầu tiên, dữ liệu được trích xuất từ External Sources – nguồn bên ngoài (giống như cách tiếp cận từ trên xuống). - Sau đó, dữ liệu đi qua Staging Area và được đưa vào Data mart thay vì Data Warehouse (như đối với cách tiếp cận từ trên xuống). Các Data mart đều được tạo trước và cung cấp khả năng báo cáo. Nó giải quyết một lĩnh vực kinh doanh duy nhất. 	<ul style="list-style-type: none"> - Báo cáo được tạo nhanh chóng. - Có thể cung cấp thêm Data mart, nhờ đó Data Warehouse có thể được mở rộng. - Tiết kiệm chi phí và thiết kế thời gian này mô hình. 	<ul style="list-style-type: none"> - Phương pháp này không mạnh bằng cách tiếp cận từ trên xuống vì chế độ hiển thị Data mart của các ô dữ liệu không nhất quán như trong cách tiếp cận trên.

Bảng 1.3 Sự khác nhau giữa cách tiếp cận Top-down và Bottom-up

1.1.5. Hạn chế của kho dữ liệu

Một số hạn chế của kho dữ liệu:

- Kho dữ liệu không phải là một lựa chọn lý tưởng cho những dữ liệu phi cấu trúc. Kho dữ liệu có thể bị lỗi thời tương đối nhanh.
- Việc thêm các nguồn dữ liệu mới sẽ mất nhiều thời gian và có chi phí cao. Đôi khi các vấn đề liên quan đến kho dữ liệu có thể không bị phát hiện trong nhiều năm.
- Kho dữ liệu là hệ thống bảo trì cao. Việc trích xuất, tải và làm sạch dữ liệu có thể tốn nhiều thời gian. Khó thực hiện thay đổi về kiểu và phạm vi dữ liệu, lược đồ nguồn dữ liệu, chỉ mục và truy vấn.
- Kho dữ liệu trông có vẻ đơn giản, nhưng thực ra, nó quá phức tạp đối với người dùng thông thường. Đôi khi người dùng kho sẽ phát triển các quy tắc kinh doanh khác nhau. Các tổ chức cần dành nhiều nguồn lực cho mục đích đào tạo và thực hiện cho người dùng cuối.
- Mặc dù đã cố gắng hết sức trong việc quản lý dự án nhưng phạm vi lưu trữ dữ liệu sẽ luôn tăng lên.
- Việc thay đổi các ràng buộc quy định có thể hạn chế khả năng kết hợp nguồn dữ liệu khác nhau. Những nguồn khác nhau này có thể bao gồm dữ liệu phi cấu trúc rất khó lưu trữ. Khi kích thước của cơ sở dữ liệu tăng lên, các ước tính về những gì tạo nên một cơ sở dữ liệu rất lớn tiếp tục phát triển. Việc xây dựng và chạy các hệ thống kho dữ liệu luôn tăng kích thước là rất phức tạp.

1.2. Tích hợp dữ liệu

1.2.1. Lý do phải tích hợp dữ liệu từ nhiều nguồn

Ngày nay, các công ty thu thập khối lượng dữ liệu khổng lồ từ nhiều nguồn khác nhau. Với việc dữ liệu ngày càng tăng theo cấp số nhân về số lượng, có nhiều định dạng khác nhau và trở nên phân tán hơn bao giờ hết. Để có dữ liệu có ý nghĩa, nó phải có thể truy cập được để phân tích nhưng hầu hết hiện nay các nguồn dữ liệu trong công ty thì được quản lý tách biệt trong các bộ phận của tổ chức ví dụ dữ liệu tài chính vẫn còn

trong bộ phận tài chính và dữ liệu tiếp thị vẫn còn trong bộ phận tiếp thị... Vì vậy, để nguồn dữ liệu từ các hệ thống quản lý khác nhau được tổng hợp thành một tập dữ liệu duy nhất chúng ta cần phải tích hợp dữ liệu. Khi một tổ chức thực hiện các bước để tích hợp dữ liệu tốt hơn, nó sẽ giảm đáng kể thời gian cần thiết để xử lý và phân tích dữ liệu.

“Tích hợp” có nghĩa là dữ liệu được lưu trữ ở các định dạng nhất quán, quy ước đặt tên, đo lường các biến, cấu trúc mã hóa, các thuộc tính vật lý của dữ liệu hoặc các ràng buộc miền (O’Leary, 1999). Tích hợp dữ liệu là phương pháp tích hợp dữ liệu từ nhiều nguồn khác nhau, không tương thích với đích kho dữ liệu thành một chế độ xem thống nhất và duy nhất để được xử lý, phân tích và chia sẻ. Tích hợp là một trong những yếu tố cốt lõi của quy trình quản lý dữ liệu tổng thể; mục tiêu chính của nó là tạo ra các tập dữ liệu hợp nhất sạch sẽ và nhất quán, đáp ứng nhu cầu thông tin của những người dùng cuối khác nhau trong một tổ chức. Đồng thời cho phép các công cụ phân tích tạo ra thông tin kinh doanh hiệu quả, có thể hành động đồng thời thúc đẩy việc đưa ra quyết định mạnh mẽ hơn, nhanh hơn. Từ đó đạt được mục tiêu cuối cùng là cung cấp cho người dùng quyền truy cập và phân phối dữ liệu nhất quán trên nhiều chủ đề và loại cấu trúc, đồng thời đáp ứng nhu cầu chia sẻ dữ liệu hiện có tiếp tục tăng lên.

Như vậy, các công ty muốn duy trì tính cạnh tranh, các công ty cần truy cập vào thông tin chính xác, phù hợp và cập nhật. Khi các hệ thống được trang bị tích hợp thời gian thực, chúng có thể nâng cao hiệu suất của chúng trên diện rộng. Tích hợp dữ liệu hỗ trợ các truy vấn trong các bộ dữ liệu khổng lồ này, hưởng lợi mọi thứ từ nghiệp vụ thông minh và phân tích dữ liệu khách hàng đến làm giàu dữ liệu và cung cấp thông tin thời gian thực. Việc thu thập dữ liệu và chuyển đổi dữ liệu thành định dạng cuối cùng, có thể sử dụng không chỉ mất ít thời gian hơn và cho phép cung cấp thông tin chi tiết có thể hành động, sự nhanh nhẹn và trí thông minh thời gian thực.

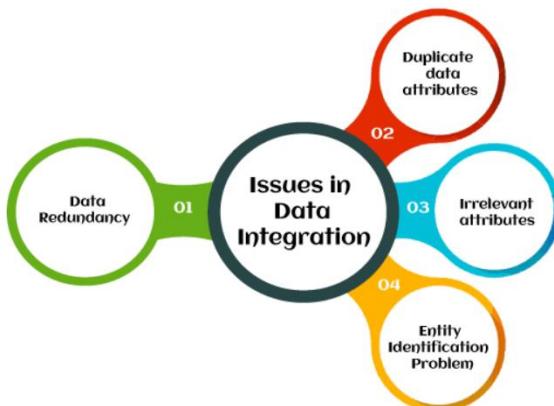
Tóm lại, khi công ty tích hợp dữ liệu thì họ sẽ nhận được rất nhiều lợi ích như:

- *Quyết định tốt hơn:* Tích hợp dữ liệu cung cấp cho các nhà phân tích dữ liệu, giám đốc điều hành công ty và nhà quản lý kinh doanh một bức tranh hoàn chỉnh về các chỉ số hiệu suất chính, khách hàng, hoạt động sản xuất và chuỗi cung ứng, nỗ lực tuân thủ quy định, rủi ro tài chính và các khía cạnh khác của quy trình kinh doanh.

Do đó, họ có sẵn thông tin phân tích tốt hơn để sử dụng như theo dõi hiệu suất kinh doanh, quản lý hoạt động và lập kế hoạch cho các chiến dịch quảng cáo và tiếp thị.

- *Tích hợp dữ liệu giúp giảm lỗi:* Khi thông tin ở nguồn cung ứng nào bị lỗi thì các bộ phận khác nhau thậm chí là các bộ phận trong các phòng ban đều có quyền truy cập vào dữ liệu của nhau, họ có thể tìm thấy lỗi nhanh hơn và có tất cả thông tin ở mọi nơi.
- *Truy cập dữ liệu có giá trị:* Những nỗ lực tích hợp dữ liệu thực sự cải thiện giá trị dữ liệu của doanh nghiệp theo thời gian. Khi dữ liệu được tích hợp vào một hệ thống tập trung, các vấn đề chất lượng dữ liệu được xác định và các cải tiến cần thiết được thực hiện, cuối cùng dẫn đến dữ liệu chính xác hơn - nền tảng cho phân tích chất lượng.
- *Cải thiện năng suất làm của nhân viên:* Với dữ liệu trải rộng trên nhiều hệ thống khác nhau, nhân viên lãng phí thời gian để tìm kiếm dữ liệu trong các hệ thống khác nhau. Việc tự động hóa các chế độ xem thống nhất giúp loại bỏ nhu cầu thu thập dữ liệu theo cách thủ công và nhân viên không còn cần phải xây dựng kết nối từ đầu bắt cứ khi nào họ cần chạy báo cáo hoặc xây dựng ứng dụng.
- *Tăng khả năng cạnh tranh:* Các công ty đảm bảo tích hợp dữ liệu trong hoạt động kinh doanh cốt lõi của họ có thể tận dụng tất cả tài sản dữ liệu và tạo ra tác động tích cực đến hiệu quả bằng cách tạo ra các sản phẩm dữ liệu phù hợp hơn.

1.2.2. Các vấn đề liên quan đến tích hợp dữ liệu



Hình 1.1 Các vấn đề trong tích hợp dữ liệu

- *Dữ liệu dư phòng*: Dữ liệu dư phòng xảy ra trong khi hợp nhất dữ liệu từ nhiều cơ sở dữ liệu. Nếu dữ liệu thừa không được loại bỏ, sẽ thu được kết quả không chính xác trong quá trình phân tích dữ liệu. Dữ liệu dư thừa xảy ra bởi những lý do sau:
 - + Nhận dạng đối tượng: Cùng một thuộc tính hoặc đối tượng có thể có các tên khác nhau trong các cơ sở dữ liệu khác nhau.
 - + Dữ liệu có thể bắt nguồn: Một thuộc tính có thể là thuộc tính “có nguồn gốc” trong một bảng khác.
- *Thuộc tính dữ liệu trùng lặp*: Cùng với việc tích hợp dữ liệu dư thừa cũng đã giải quyết các bộ giá trị trùng lặp. Các bộ giá trị trùng lặp có thể xuất hiện trong dữ liệu kết quả nếu bảng không chuẩn hóa đã được sử dụng làm nguồn để tích hợp dữ liệu.
- *Thuộc tính không liên quan*: Một số thuộc tính trong dữ liệu không quan trọng và chúng không được xem xét khi thực hiện các tác vụ khai thác dữ liệu. Không có ích lợi gì khi có các thuộc tính không liên quan như vậy trong dữ liệu.
- *Vấn đề nhận dạng thực thể*: Sự cố nhận dạng thực thể xảy ra trong quá trình tích hợp dữ liệu. Trong khi tích hợp dữ liệu từ nhiều tài nguyên, một số tài nguyên dữ liệu phù hợp với nhau, trớn nên có giá trị nếu chúng được tích hợp. Vấn đề nhận dạng thực thể giúp phát hiện và giải quyết các xung đột về giá trị dữ liệu. Dữ liệu thường được thu thập từ nhiều tài nguyên vào một kho lưu trữ nhất quán và nó có thể có các kích thước và kiểu dữ liệu khác nhau. Có các cách biểu diễn dữ liệu khác nhau và các quy mô dữ liệu khác nhau. Các vấn đề nhận dạng thực thể có thể xảy ra trong cả tích hợp cơ sở dữ liệu ảo và thực tế.
 - + Tích hợp ảo: Một cơ sở dữ liệu được tích hợp ảo được tạo trên đầu các cơ sở dữ liệu thành phần, thường sử dụng một mô hình dữ liệu chung và lược đồ tích hợp. Các thành phần giữ lại danh tính và cách sử dụng của chúng. Nỗ lực trong các cơ sở dữ liệu tự trị liên hợp là theo hướng này.
 - + Tích hợp thực tế: Một cơ sở dữ liệu tích hợp thực sự được tạo ra từ các cơ sở dữ liệu thành phần. Cơ sở dữ liệu ban đầu bị loại bỏ và các ứng dụng được chuyển sang cơ sở dữ liệu tích hợp mới.

Trong một ngữ cảnh cơ sở dữ liệu duy nhất, thường là trường hợp một cá thể đối tượng có thể tạo mô hình duy nhất cho một thực thể trong thế giới thực. Thuộc tính này không áp dụng cho nhiều cơ sở dữ liệu tự trị, do đó nảy sinh vấn đề nhận dạng thực thể.

Cơ sở dữ liệu đã có từ trước trong hầu hết các tổ chức được xác định và điền bởi những người khác nhau vào những thời điểm khác nhau để đáp ứng các yêu cầu khác nhau của tổ chức hoặc người dùng cuối. Việc phát triển cơ sở dữ liệu độc lập như vậy thường dẫn đến hai cơ sở dữ liệu nắm bắt các phần của cùng một miền trong thế giới thực. Thông thường, khi cần cung cấp quyền truy cập tích hợp vào các cơ sở dữ liệu liên quan này, việc liên hệ các biểu diễn của cùng một thực thể trong thế giới thực từ hai cơ sở dữ liệu thường khó, nếu không muốn nói là không thể, nếu không chỉ định thông tin ngữ nghĩa bổ sung để giải quyết sự mơ hồ này.

- *Phát hiện và giải quyết xung đột dữ liệu:* Xung đột dữ liệu có nghĩa là dữ liệu được hợp nhất từ các nguồn khác nhau không khớp. Giống như các giá trị thuộc tính có thể khác nhau trong các tập dữ liệu khác nhau. Sự khác biệt có thể do chúng được biểu diễn khác nhau trong các tập dữ liệu khác nhau. Giả sử giá phòng khách sạn có thể được thể hiện bằng các đơn vị tiền tệ khác nhau ở các thành phố khác nhau. Loại vấn đề này được phát hiện và giải quyết trong quá trình tích hợp dữ liệu.

1.3. Quản trị và quản lý dữ liệu

1.3.1. Phân biệt quản trị và quản lý dữ liệu

Quản trị dữ liệu (Data Governance) là một tập hợp các nguyên tắc cùng các hoạt động cần thực hiện nhằm đảm bảo chất lượng dữ liệu trong suốt vòng đời hoàn chỉnh của nó. Hệ thống các quy định trong quản trị dữ liệu giúp xác định chủ sở hữu dữ liệu, người/ tổ chức có quyền kiểm soát đối với các tài sản dữ liệu và cách các tài sản đó có thể được sử dụng.

Theo quan điểm của Viện quản trị dữ liệu tại Mỹ (The Data Governance Institute)

Quản lý dữ liệu là một thuật ngữ bao quát, mô tả các quy trình được sử dụng để lập kế hoạch, chỉ định, kích hoạt, khởi tạo, thu nhập, duy trì, sử dụng, lưu trữ, truy xuất, kiểm soát và xóa/hủy dữ liệu.

Theo DAMA, 2017

- ⇒ Quản trị dữ liệu là việc thực thi quyền lực, kiểm soát và ra quyết định chung (lập kế hoạch, giám sát và thi hành). Những sáng kiến quản trị dữ liệu cùng các nền tảng để phát triển các giao thức và thủ tục để qua đó quản lý dữ liệu hợp lý. Mặt khác, quản lý dữ liệu là quá trình đưa chính sách quản trị vào hoạt động. Quản trị cung cấp một khuôn khổ để các cơ quan, tổ chức, doanh nghiệp có thể xác định các khu vực để quản lý (như bảo mật, cơ sở dữ liệu và kiểm soát dữ liệu). Hay nói cách khác Quản trị dữ liệu là việc thiết lập lý do tại sao và ai cho khả năng truy cập và kiểm soát dữ liệu, trong khi Quản lý dữ liệu thiết lập nơi truy cập và cách thức truy cập dữ liệu.
- ⇒ Khi nói đến việc thực hiện một kế hoạch quản lý dữ liệu, quản trị dữ liệu cung cấp hướng cần thiết.

Quản trị dữ liệu	Quản lý dữ liệu
Liên quan đến các chính sách quy định và thủ tục để quản lý chất lượng dữ liệu.	Đề cập đến việc dữ liệu được quản lý.
Đề cập đến việc áp dụng kiến thức, phát triển thủ tục và hình thành lý thuyết.	Đề cập đến việc thu thập, tổ chức, bảo vệ và xử lý, sắp xếp và bảo mật dữ liệu.
Nó nhằm mục đích đảm bảo tính chính xác và toàn vẹn dữ liệu được lưu trữ.	Nó nhằm mục đích cải thiện chất lượng tổng thể và giá trị tài chính.
Tập trung vào triết học và kinh doanh.	Tập trung vào logic và công nghệ
Đó là một chiến lược để có được liệu chất lượng cao.	Đó là phương pháp để sắp xếp các sự kiện một cách hợp lý.

Bảng 1.4 So sánh Quản trị dữ liệu và Quản lý dữ liệu

1.3.2. Lý do phải quản lý dữ liệu chủ

Quản lý dữ liệu chủ là tập hợp các hoạt động quản lý “dữ liệu tốt nhất”, mà có thể điều phối các bên liên quan chính, những người tham gia và khách hàng doanh nghiệp trong việc kết hợp với các ứng dụng kinh doanh, các phương pháp quản lý thông tin và công cụ quản lý dữ liệu để thực thi các chính sách, thủ tục, dịch vụ và cơ sở hạ tầng hỗ trợ cho việc nắm bắt, tích hợp và sau đó chia sẻ việc sử dụng dữ liệu chủ, đảm bảo dữ liệu này là chính xác, kịp thời, nhất quán. (David Loshin, 2009).

Cần phải quản lý dữ liệu chủ bởi quản lý dữ liệu chủ thực sự quan trọng bởi những lý do sau:

- Tích hợp thông tin trên nhiều nền tảng

Có dữ liệu chính xác và nhất quán là rất quan trọng để cung cấp cho khách hàng để cung cấp cho khách hàng của bạn trải nghiệm mua sắm tùy chỉnh và tạo sự khác biệt cho doanh nghiệp của bạn trên thị trường. Quản lý dữ liệu chủ tạo ra một tập tin tổng thể về thông tin sản phẩm chính xác, chất lượng cao để phân phối thích hợp thông qua tất cả các con đường bán hàng.

Trong hầu hết các tổ chức không có quản lý dữ liệu chủ, thông tin doanh nghiệp thường ở các silo khác nhau, nơi mỗi bộ phận hoặc đơn vị kinh doanh quản lý dữ liệu của mình. Đôi khi, dữ liệu có sẵn ở nhiều định dạng. Như vậy, việc dư thừa dữ liệu là điều khó tránh khỏi. Quản lý dữ liệu chủ đạt điểm ở đây bằng cách cân bằng cách tập hợp tất cả dữ liệu từ các nền tảng khác nhau như trực tuyến, vật lý, đám mây và những nơi khác. Bằng cách này, người sử dụng thông tin sẽ không còn phải sử dụng thông tin sẽ không còn phải xem thông tin từ nhiều kênh để tạo báo cáo vì tất cả dữ liệu sẽ được trình bày trong một chế độ xem thống nhất duy nhất với các chi tiết cụ thể và mạch lạc hoàn toàn.

- Tạo sự hiểu biết tốt hơn về khách hàng và doanh nghiệp

Thông qua quản lý dữ liệu chủ, doanh nghiệp tổ chức có thể đồng bộ hóa tất cả dữ liệu khách hàng trên nhiều kênh và chuỗi thông tin để duy trì một chế độ xem.

Chương trình giúp bạn có cái nhìn tổng hợp về vật liệu, khách hàng, nhà cung cấp và các tập dữ liệu khác. Khi chúng ta có sẵn tất cả thông tin doanh nghiệp, chúng ta có thể tiếp cận và ra quyết định một cách kịp thời.

- **Tăng độ tin cậy của dữ liệu**

Các doanh nghiệp dựa vào dữ liệu để tối đa hóa hiệu quả hoạt động của mình. Mỗi khi tạo báo cáo, chúng ta có nhiều khả năng dựa vào dữ liệu trong hệ thống của mình hơn. Nếu chúng ta đang làm việc với dữ liệu không đầy đủ, các báo cáo của bạn sẽ không phản ánh thực tế của doanh nghiệp và chúng ta không thể nhận được toàn bộ lợi ích. Tệ hơn nữa sẽ ảnh hưởng đến quá trình ra quyết định của doanh nghiệp tổ chức chúng ta. Việc quản lý dữ liệu chủ thành công giúp đơn giản hóa việc quản trị dữ liệu. Quản lý dữ liệu chủ loại bỏ những nhầm lẫn và tạo ra một quy trình làm việc nhất quán và đơn giản hóa.

- **Tính linh hoạt và chỉnh sửa dữ liệu**

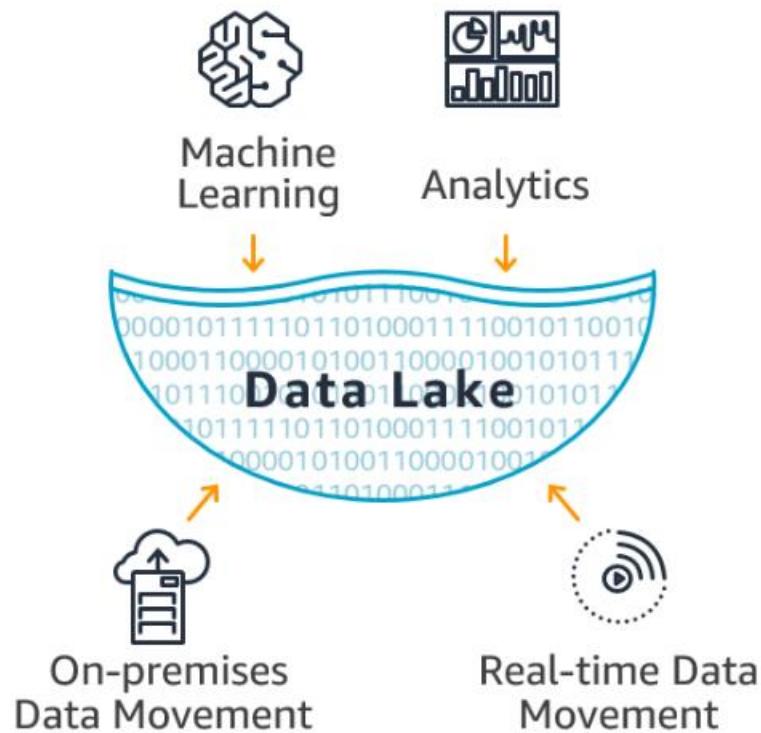
Một thách thức của việc giữ nhiều kênh dữ liệu là các thay đổi được thực hiện đối với dữ liệu ở một vị trí sẽ không được phản ánh trên tất cả các kênh. Cách tiếp cận này có nhiều khả năng dẫn đến sai lệch dữ liệu. Quản lý dữ liệu chủ cho phép thực hiện các chỉnh sửa tổng thể theo cách mà nó cập nhật các bản ghi bị ảnh hưởng ở tất cả các vị trí do đó đảm bảo tính nhất quán của dữ liệu. Bên cạnh đó, tính linh hoạt của dữ liệu của quản lý dữ liệu chủ cao hơn so với các chiến lược quản lý dữ liệu khác.

1.4. Tổng quan về hồ dữ liệu (Data Lake)

1.4.1. Khái niệm hồ dữ liệu

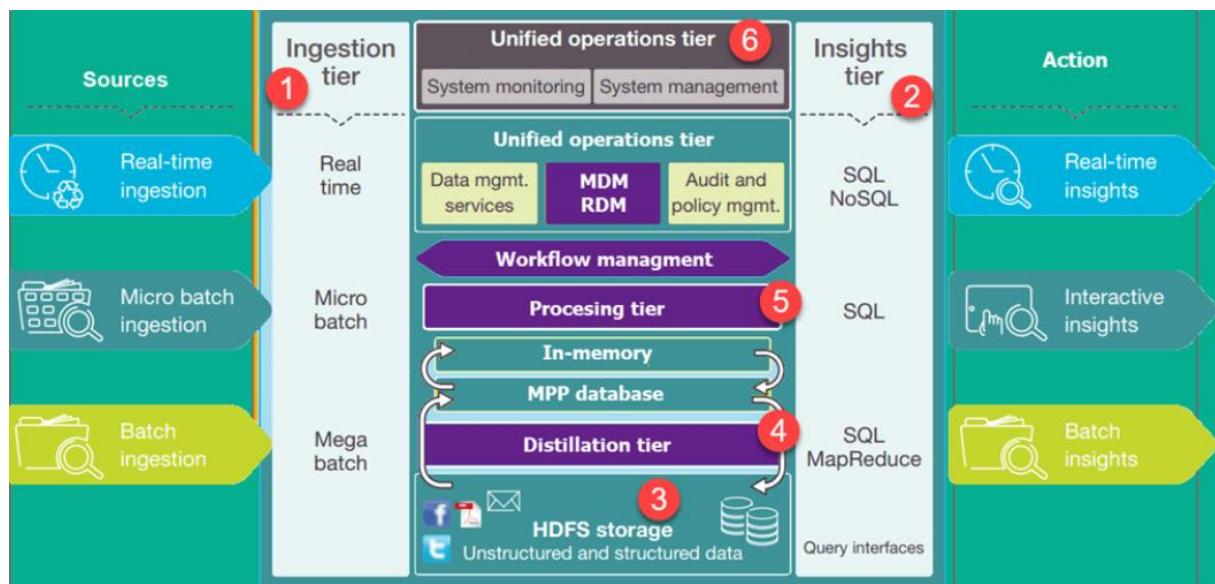
Một hồ dữ liệu cho phép bạn lưu trữ tất cả dữ liệu có cấu trúc và phi cấu trúc của mình, trong một kho lưu trữ tập trung và ở bất kỳ quy mô nào. Với hồ dữ liệu, bạn có thể lưu trữ dữ liệu của mình nguyên trạng mà không cần phải cấu trúc dữ liệu trước, dựa trên các câu hỏi tiềm năng mà bạn có thể có trong tương lai. Các hồ dữ liệu cũng cho phép bạn chạy các loại phân tích khác nhau trên dữ liệu của mình như truy vấn SQL,

phân tích dữ liệu lớn, tìm kiếm toàn văn, phân tích thời gian thực và học máy để đưa ra các quyết định tốt hơn.



Hình 1.2 Hồ dữ liệu (Nguồn: Amazon Web Services)

1.4.2. Kiến trúc Data Lake



Hình 1.3 Kiến trúc Data Lake (Nguồn: guru99.com)

Sơ đồ mô phỏng kiến trúc Data Lake trên cho chúng ta thấy các cấp độ thể hiện dữ liệu ở các cấp có các trạng thái khác nhau, ví dụ như ở cấp thấp hơn thì sẽ thể hiện dữ liệu hầu như ở trạng thái rest (trạng thái nghỉ) trong khi các cấp trên hiển thị dữ liệu giao dịch theo real-time (thời gian thực) và luồng dữ liệu này qua hệ thống không có hoặc có độ trễ ít.

Các tầng trong sơ đồ:

- 1. Ingestion Tier (Tầng nạp dữ liệu):** Các lớp ở phía bên trái mô tả các nguồn dữ liệu. Dữ liệu có thể được tải vào hồ dữ liệu theo lô hoặc trong thời gian thực
- 2. Insights Tier (Tầng khai phá):** Các cấp ở bên phải đại diện cho phía nghiên cứu, nơi những thông tin chi tiết từ hệ thống được sử dụng. Các truy vấn SQL, NoSQL hoặc thậm chí excel có thể được sử dụng để phân tích dữ liệu.
- 3. HDFS storage (Tầng lưu trữ):** Là một giải pháp hiệu quả về chi phí cho cả dữ liệu có cấu trúc và không có cấu trúc. Nó là bối cảnh cho tất cả dữ liệu ở trạng thái nghỉ trong hệ thống.
- 4. Distillation tier (Tầng tiền xử lý):** Lấy dữ liệu từ bộ nhớ lưu trữ và chuyển đổi nó thành dữ liệu có cấu trúc để phân tích dễ dàng hơn.
- 5. Processing tier (Tầng xử lý):** Chạy các thuật toán phân tích và truy vấn người dùng với thời gian thực khác nhau, tương tác, hàng loạt để tạo dữ liệu có cấu trúc để phân tích dễ dàng hơn.
- 6. Unified operations tier (Tầng giám sát, vận hành):** Chi phối quản lý và giám sát hệ thống. Nó bao gồm kiểm toán và quản lý thành thạo, quản lý dữ liệu, quản lý quy trình làm việc.

1.4.3. Ưu điểm và hạn chế của Data Lake trong quá trình sử dụng

1.4.3.1. Ưu điểm

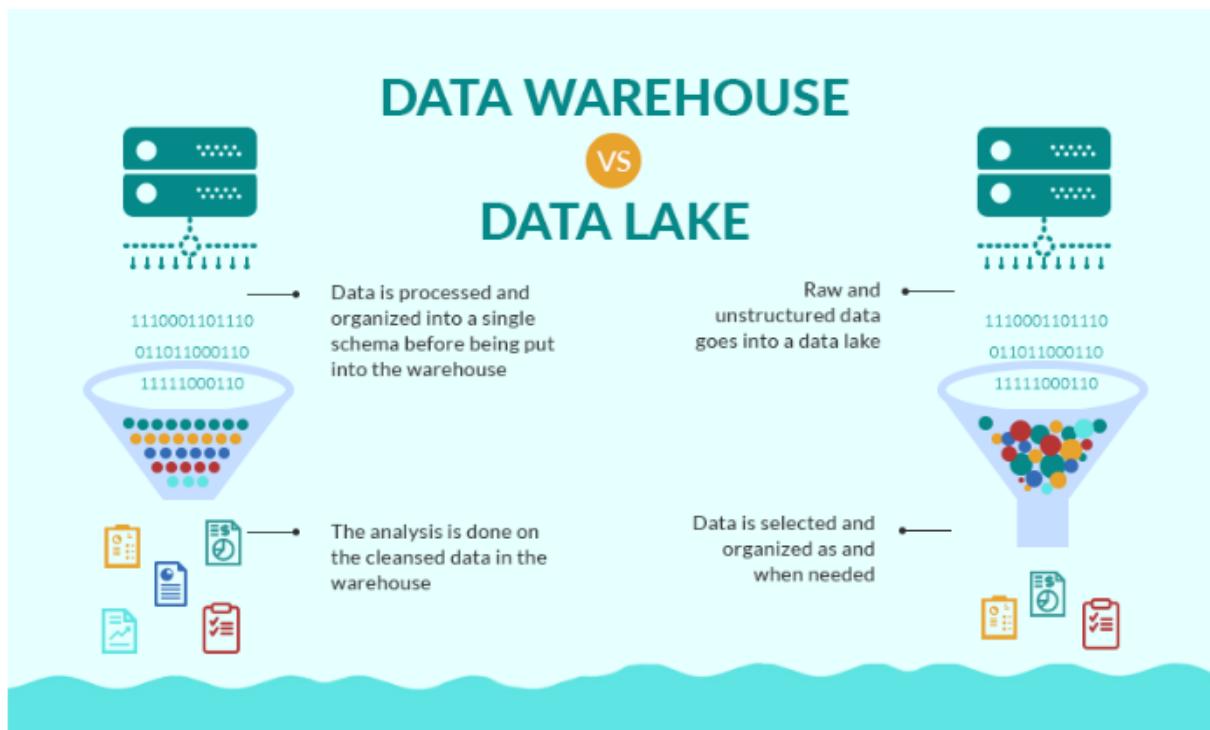
- Hồ dữ liệu là một nền tảng lưu trữ linh hoạt có thể dễ dàng định cấu hình cho bất kỳ mô hình, cấu trúc, ứng dụng hoặc truy vấn dữ liệu nhất định nào. Tính linh hoạt của hồ dữ liệu cho phép nhiều phương pháp phân tích nâng cao và tiên tiến để giải thích dữ liệu.

- Là một lược đồ khi đọc làm cho một hồ dữ liệu có thể mở rộng và linh hoạt.
- Các hồ dữ liệu hỗ trợ các truy vấn yêu cầu phân tích sâu bằng cách khám phá thông tin từ nguồn của nó đến các truy vấn yêu cầu một báo cáo đơn giản với dữ liệu tóm tắt. Tất cả các loại người dùng được phục vụ cho.
- Hầu hết các ứng dụng phần mềm của data lake đều là mã nguồn mở và có thể được cài đặt bằng phần cứng giá rẻ.
- Việc phát triển lược đồ được hoãn lại cho đến khi một tổ chức tìm thấy một trường hợp nghiệp vụ cho dữ liệu. Do đó, không lãng phí thời gian và chi phí cho việc phát triển lược đồ.
- Các hồ dữ liệu cung cấp sự tập trung của các nguồn dữ liệu khác nhau.
- Chúng cung cấp giá trị cho tất cả các loại dữ liệu cũng như chi phí sở hữu lâu dài.
- Các hồ dữ liệu dựa trên đám mây dễ triển khai hơn và nhanh hơn, tiết kiệm chi phí với mô hình trả tiền khi sử dụng và dễ dàng mở rộng quy mô hơn khi có nhu cầu. Nó cũng tiết kiệm không gian và chi phí bất động sản.

1.4.3.2. Hạn chế

- Các hồ dữ liệu có nguy cơ mất liên quan và trở thành đám dữ liệu theo thời gian nếu chúng không được quản lý thích hợp.
- Chi phí lưu trữ và xử lý có thể tăng lên khi nhiều dữ liệu được thêm vào hồ.
- Các hồ dữ liệu tại chỗ phải đổi mặt với những thách thức như hạn chế về không gian, thiết lập phần cứng và trung tâm dữ liệu, khả năng mở rộng lưu trữ, chi phí và ngân sách tài nguyên.
- Không có cách nào để có được thông tin chi tiết từ những người khác đã làm việc với dữ liệu vì không có tài khoản nào về nguồn gốc của các phát hiện của các nhà phân tích trước đó.
- Rủi ro lớn nhất của các hồ dữ liệu là bảo mật và kiểm soát truy cập. Đôi khi dữ liệu có thể được đưa vào hồ mà không cần bất kỳ sự giám sát nào, vì một số dữ liệu có thể có quyền riêng tư và nhu cầu pháp lý.

1.4.4. So sánh Data Warehouse và Data Lake



Hình 1.4 Sự khác nhau giữa Data Warehouse và Data Lake (Nguồn: grazitti.com)

Giống: Cả hai đều là kho dữ liệu phục vụ cùng một mục đích chung và mục tiêu lưu trữ dữ liệu tổ chức để hỗ trợ việc ra quyết định.

Khác:

	Data Lake	Data Warehouse
Data	Các hồ dữ liệu lưu trữ mọi thứ.	Kho dữ liệu chỉ tập trung vào Quy trình nghiệp vụ.
Xử lý	Dữ liệu chủ yếu chưa được xử lý	Dữ liệu được xử lý cao.
Loại dữ liệu	Data lake bao gồm tất cả các loại dữ liệu và cấu trúc, bán cấu trúc và không cấu trúc ở dạng ban đầu của chúng từ các hệ thống nguồn.	Data Warehouse bao gồm các thông tin có cấu trúc chúng được sắp xếp trong các lược đồ và được xác định cho mục đích kho dữ liệu.

Data Timeline	Data Lake có thể giữ lại tất cả dữ liệu. Điều này không chỉ bao gồm dữ liệu đang được sử dụng mà còn bao gồm dữ liệu mà nó có thể sử dụng trong tương lai. Ngoài ra, dữ liệu được lưu giữ mọi lúc để quay ngược thời gian và thực hiện phân tích, xử lý.	Trong quá trình phát triển kho dữ liệu, thời gian chủ yếu được dành cho việc phân tích các nguồn dữ liệu khác nhau.
Đối tượng sử dụng	Hồ dữ liệu lý tưởng cho những người dùng phân tích sau: nhà khoa học dữ liệu, những người cần các công cụ phân tích tiên tiến với các khả năng như mô hình dự đoán và phân tích thống kê.	Kho dữ liệu lý tưởng cho người dùng vận hành (operational users) vì được cấu trúc tốt, dễ sử dụng và dễ hiểu.
Chi phí lưu trữ	Việc lưu trữ dữ liệu trong công nghệ dữ liệu lớn tương đối rẻ so với việc lưu trữ dữ liệu trong kho dữ liệu.	Lưu trữ dữ liệu trong Kho dữ liệu tốn kém hơn và tốn thời gian.
Vị trí của lược đồ (schema)	Lược đồ khi đọc (không có lược đồ xác định trước)	Lược đồ khi ghi (lược đồ xác định trước)
Xử lý dữ liệu	Data Lakes sử dụng quy trình ELT (Extract Load Transform).	Kho dữ liệu sử dụng quy trình ETL (Extract Transform Load) truyền thống.
Mức độ chi tiết của dữ liệu	Dữ liệu ở mức độ chi tiết hoặc chi tiết thấp.	Dữ liệu ở mức độ chi tiết tóm tắt hoặc tổng hợp.
Bảo mật	Cung cấp khả năng kiểm soát ít hơn.	Cho phép kiểm soát dữ liệu tốt hơn.

Bảng 1.5 Mô tả sự khác nhau giữa Data Lake và Data warehouse (Nguồn: guru99.com)

CHƯƠNG 2. TỔNG QUAN VỀ CÔNG TY

2.1. Giới thiệu về công ty

- *Tên công ty:* CÔNG TY CỔ PHẦN CÔNG NGHỆ GIÁO DỤC EDTE.
- *Tên quốc tế:* EDTE EDUCATION TECHNOLOGY CORPORATION.
- *Tên viết tắt:* EDTE., CORP.
- *Mã số thuế:* 0107122138.
- *Địa chỉ:* C17+18, Lô 20, Khu đô thị mới Định Công, Phường Định Công, Quận Hoàng Mai, Thành phố Hà Nội, Việt Nam.
- *Văn phòng đại diện:* 35 Tô Vĩnh Diện, Khương Trung, Thanh Xuân, Hà Nội.
- Công ty có Giấy chứng nhận đăng ký kinh doanh từ ngày 17 tháng 11 năm 2015.
- *Loại hình doanh nghiệp:* Công ty cổ phần ngoài nhà nước.
- *Ngành nghề kinh doanh:* Kinh doanh và giảng dạy trực tuyến các khóa học qua video, khóa học giao tiếp Tiếng Nhật 1:1 cho người đi làm, khóa học luyện thi JLPT, khóa học giao tiếp tiếng Trung Quốc.
- Sơ đồ công ty:



Hình 2.1 Sơ đồ cơ cấu tổ chức công ty cổ phần giáo dục EDTE

2.2. Hiện trạng

Trên thực tế, mọi hoạt động kinh doanh, vận hành lớp học, quản lý giáo viên, quản lý doanh thu, quản lý học viên, tính lương... của trung tâm đều đang được ghi chép lại ở Excel làm cho việc quản lý gặp nhiều khó khăn. Việc làm này chủ yếu là thủ công nên tình trạng chồng chéo dữ liệu, không đồng nhất làm mất rất nhiều thời gian quản lý.

Dữ liệu giữa các phòng ban không đồng nhất chẳng hạn như:

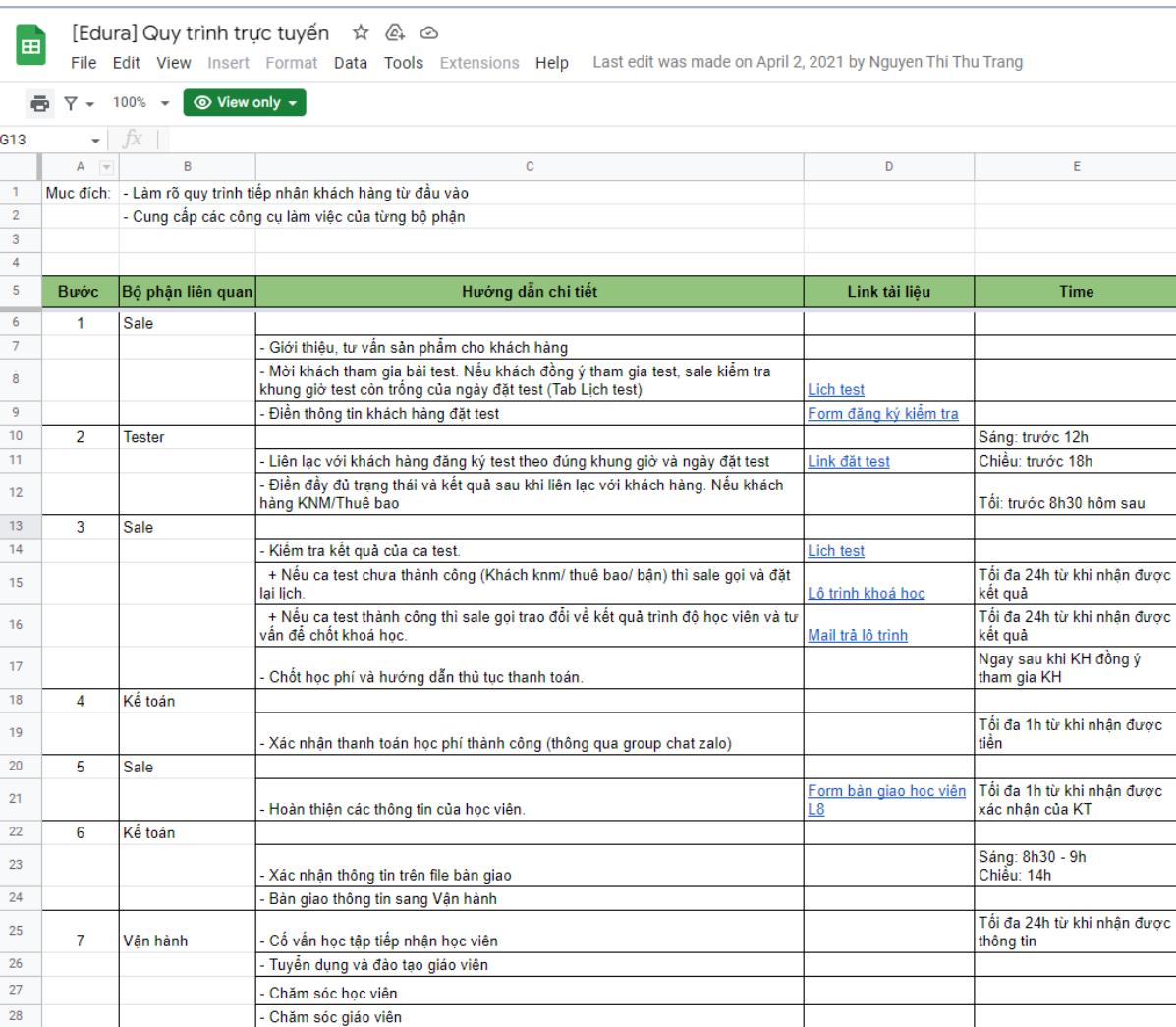
- Khi phòng Marketing lấy data khách hàng, họ tên khách hàng chỉ là tên các tài khoản mạng xã hội, nhưng khi khách hàng đã đăng ký và trở thành học viên của trung tâm thì Phòng Đào Tạo (Phòng BO) sẽ lấy họ tên khai sinh của học viên để tiến hành lập hồ sơ tham gia học. Việc này đã tạo thông tin không đồng bộ giữa các phòng ban.
- Học viên nộp hoàn thiện học phí được phòng Sale cập nhật lên file nhưng phòng Đào Tạo chỉ nhận được lời thông báo mới có thể nắm thông tin để điều chỉnh số buổi học của học viên một cách thủ công. (Học viên có thể đóng trước 60% học phí và nhận được 10 buổi (tương ứng 30% số buổi học của khóa))
- Học viên chủ yếu là là người học tập công tác tại Nhật, thực tập sinh, du học sinh sắp qua Nhật vì vậy mỗi học viên sẽ có nhiều địa chỉ khác nhau: có thể là địa chỉ thường trú, địa chỉ tạm trú, nguyên quán, hoặc địa chỉ hiện tại ở Nhật, tùy vào sự cung cấp thông tin của học viên cho từng phòng ban.
- Phòng Marketing, Phòng Sale, Phòng đào tạo liên lạc với khách hàng (học viên) qua các kênh khác nhau: một học viên có thể liên lạc với Phòng Marketing qua Facebook, liên lạc với phòng Sale qua số điện thoại và liên lạc với Phòng đào tạo qua Zalo/Line,

Để khắc phục những bất cập này, nhóm xin đề xuất xây dựng một hệ thống kho dữ liệu riêng nhằm cải thiện việc kiểm soát dữ liệu và tạo trải nghiệm tốt nhất cho học viên của trung tâm và để đưa ra các báo cáo, phân tích, dự đoán khách hàng một cách tốt nhất. Từ đó, phục vụ cho việc ra quyết định.

2.3. Quy trình phân tích

Công ty Edura kinh doanh và giảng dạy các khóa học giao tiếp tiếng Nhật thông qua một hệ thống riêng. Quy trình kinh doanh và quy trình quản lý học tập của học viên là hai quy trình trọng tâm đối với sự phát triển của công ty.

Mỗi phòng được xác định với một chức năng cụ thể. Quy trình kinh doanh sẽ được vận hành bởi phòng Sales và quy trình quản lý học tập sẽ được thực hiện chủ yếu bởi Phòng đào tạo.



The screenshot shows a Google Sheets document titled "[Edura] Quy trình trực tuyến". The menu bar includes File, Edit, View, Insert, Format, Data, Tools, Extensions, Help. A note at the top right says "Last edit was made on April 2, 2021 by Nguyen Thi Thu Trang". The view is set to "View only". The sheet is titled "G13" and contains two tables.

Table 1: Mục đích

	A	B	C	D	E
1	Mục đích:	<ul style="list-style-type: none"> - Làm rõ quy trình tiếp nhận khách hàng từ đầu vào - Cung cấp các công cụ làm việc của từng bộ phận 			
2					
3					
4					

Table 2: Quy trình làm việc

	Bước	Bộ phận liên quan	Hướng dẫn chi tiết	Link tài liệu	Time
6	1	Sale	<ul style="list-style-type: none"> - Giới thiệu, tư vấn sản phẩm cho khách hàng - Mời khách tham gia bài test. Nếu khách đồng ý tham gia test, sale kiểm tra khung giờ test còn trống của ngày đặt test (Tab Lịch test) - Điện thông tin khách hàng đặt test 		
7					
8				Lịch test	
9				Form đăng ký kiểm tra	
10	2	Tester	<ul style="list-style-type: none"> - Liên lạc với khách hàng đăng ký test theo đúng khung giờ và ngày đặt test - Điện đầy đủ trạng thái và kết quả sau khi liên lạc với khách hàng. Nếu khách hàng KNM/Thuê bao 	Link đặt test Sáng: trước 12h Chiều: trước 18h	
11					
12					Tối: trước 8h30 hôm sau
13	3	Sale	<ul style="list-style-type: none"> - Kiểm tra kết quả của ca test. + Nếu ca test chưa thành công (Khách kmn/ thuê bao/ bđm) thì sale gọi và đặt lại lịch. + Nếu ca test thành công thì sale gọi trao đổi về kết quả trình độ học viên và tư vấn để chốt khóa học. - Chốt học phí và hướng dẫn thủ tục thanh toán. 	Lịch test Lô trình khóa học Mail trả lô trình Ngay sau khi KH đồng ý tham gia KH	
14					
15					Tối đa 24h từ khi nhận được kết quả
16					Tối đa 24h từ khi nhận được kết quả
17					
18	4	Kế toán			
19			<ul style="list-style-type: none"> - Xác nhận thanh toán học phí thành công (thông qua group chat zalo) 		Tối đa 1h từ khi nhận được tiền
20	5	Sale			
21			<ul style="list-style-type: none"> - Hoàn thiện các thông tin của học viên. 	Form bàn giao học viên L8	Tối đa 1h từ khi nhận được xác nhận của KT
22	6	Kế toán			
23			<ul style="list-style-type: none"> - Xác nhận thông tin trên file bàn giao - Bàn giao thông tin sang Vận hành 		Sáng: 8h30 - 9h Chiều: 14h
24					
25	7	Vận hành	<ul style="list-style-type: none"> - Cố vấn học tập tiếp nhận học viên - Tuyển dụng và đào tạo giáo viên - Chăm sóc học viên - Chăm sóc giáo viên 		Tối đa 24h từ khi nhận được thông tin
26					
27					
28					

Hình 2.2 Quy trình làm việc giữa các phòng ban tại công ty Edura (Nguồn: Công ty Edura)

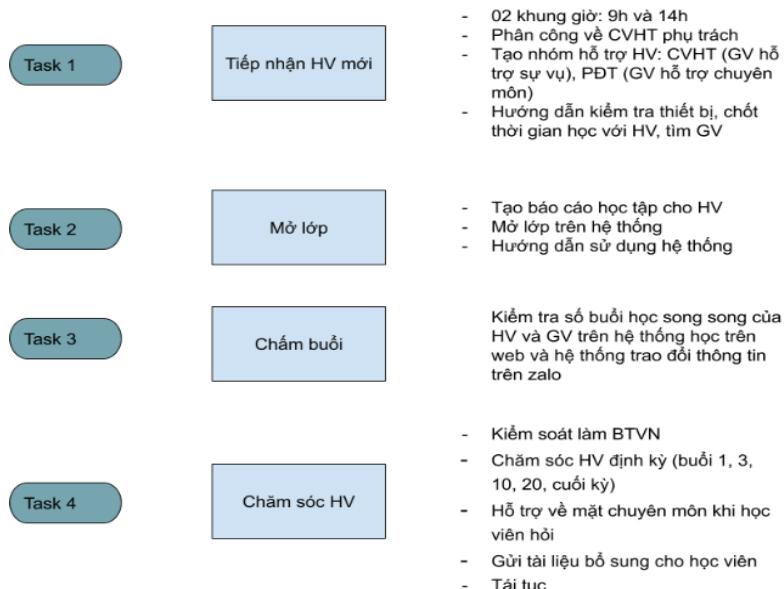
Thời gian	Chức năng Sale	Chức năng PDT
Trước khi bàn giao L8	Tư vấn đầy đủ và chính xác về chương trình cho học viên.	Hỗ trợ tìm kiếm và trả lời các lớp học có lịch đặc biệt (Chỉ áp dụng đối với GV Việt Nam)
	Gửi đầy đủ "Bản cam kết và nội quy" cho HV trước khi vào học.	
	Xác nhận rõ với học viên về vấn đề thanh toán học phí (số tiền đã đóng, số buổi được học, thời hạn hoàn thiện nốt học phí)	
Sau khi bàn giao L8	Nếu học viên có lịch học đặc biệt, cần email cho phongdaotao@edura.vn và cc cho quản lý để được hỗ trợ tìm giáo viên. Không chấp nhận bất cứ hình thức tư sắp xếp nào không thông qua email.	
	Chịu trách nhiệm tất cả các vấn đề phát sinh từ học viên khi có điểm nóng về vấn đề tu vấn sai thông tin khóa học, giáo viên, học phí.	Tiếp nhận, sắp xếp lớp học và chăm sóc học viên theo đúng quy định
	Tiếp nhận thông tin khi học viên đang học phản hồi về lịch học, giáo viên, lô trình...và chuyển sang cho PDT xử lý. Không tự ý làm việc trực tiếp với học viên không thông qua PDT	Chịu trách nhiệm xử lý các vấn đề phát sinh/ điểm nóng của học viên trong quá trình học liên quan đến giáo trình, giáo viên, hệ thống (không phải do tu vấn sai)
Trong vòng 45 ngày kể từ ngày học viên đóng học phí		Tiếp nhận thông tin từ sale về phản hồi của học viên trong quá trình học. Xác nhận lại với sale về việc tiếp nhận thông tin và phương án xử lý
	Được quyền nâng cấp gói học cho học viên, đóng học phí lần 2, 3...	Hỗ trợ sale cung cấp các thông tin liên quan đến quá trình học của những học viên nâng cấp gói, đóng học phí lần 2, 3...
Sau 45 ngày kể từ ngày đóng học phí		Được quyền cho học viên gia hạn, xử lý các khoản đóng học phí lần 2, 3 quá hạn

Hình 2.3 Chức năng của mỗi phòng (Nguồn: Công ty Edura)

2.3.1. Quy trình quản lý học tập

Học viên sẽ được quản lý bởi phòng đào tạo, mỗi lớp học sẽ dạy sẽ gồm 1 giáo viên và 1 học viên (có thể học nhóm 1 giáo viên - 4 học viên).

Quy trình quản lý học tập bao gồm các quy trình nhỏ như quản lý học viên, quản lý giáo viên, vận hành lớp học.



Hình 2.4 Quy trình quản lý học viên (Nguồn: Công ty Edura)

Edura_bộ phận vận hành							
	A	B	C	D	E	F	G
1	Các đầu công việc của Hỗ trợ vận hành						
2	Số thứ tự	Nội dung công việc	Chi tiết công việc	Mục đích	Các file/Hướng dẫn có liên quan	Tần suất	KPI
3	1	Check Bài Tập Về Nhà	- Kiểm tra BTVN của HV trên hệ thống, điền vào file Chăm sóc về kết quả check - Nhắc nhở GV chấm bài - Nhắc nhở HV bổ sung những bài chưa làm	- Kiểm tra xem Học viên có làm BTVN đầy đủ không theo quy trình	Vào mục Báo cáo thống kê -> Báo cáo kết quả bài tập về	3 buổi/lần	100% Học viên và Giáo viên
4	2	Check Báo Cáo Học Tập	- Kiểm tra xem GV sau mỗi 10 buổi đã hoàn thiện BCHT cho HV chưa	- Theo dõi quá trình học của HV	Phản BCHT trong sheet Infor trên file GV	10 buổi/lần	100% GV phải làm BCHT cho HV
5	3	Check thông báo xin nghỉ của GV và HV	Tiếp nhận thông báo xin nghỉ từ GV và HV, phản hồi lại và trao đổi về lịch học bù (nếu có)	Nhắm được số buổi nghỉ của GV và HV	Điều chỉnh nghỉ trên file [2022CVHT01] và [2022CVHT02]	Theo ca	
6	4	Check buổi	- Check buổi đối với số buổi học còn lại của HV + Tổng cộng có bao nhiêu ca HV và GV xin nghỉ	Phản ánh được đúng số buổi HV đã học	Check buổi trên file [2022CVHT01] và [2022CVHT02]	Hàng ngày	Phản ánh được đúng số buổi HV đã học
7	5	Tiếp nhận HV mới	- Gửi 3 mail có trong temple cho HV - Kết bạn với HV mới và tạo nhóm - Gửi thông tin ban đầu về số buổi học và giờ học cho HV	Tiếp nhận HV mới	HV mới được bàn giao trên file [2022CVHT01] và [2022CVHT02]	Theo ca (09h30 và 14h30 hàng ngày)	
8							
9							
10			Đối với còn 60 buổi, xem HV có cần lên khóa mới hay chưa + Đối với còn 30 buổi: xem HV có học được với GV cũ tiếp không hay chuyển lớp sang GV mới/GV Nhật và mở lớp mới + Đối với còn 3 buổi: Nhắc HV và GV về số buổi học còn lại của HV + Đối với HV còn 0 buổi: xóa lệnh, đổi trang thái Kết thúc, khóa lớp và gửi BCHT cho HV				

Hình 2.5 Quy trình vận hành lớp học (Nguồn: Công ty Edura)

Thời gian	Công việc phải làm	File vận hành	Xử lý
8h30 - 10h00	- Check file "Nhật ký trực ca tối" và giải quyết những vấn đề tồn đọng - Check Zalo để hỗ trợ học viên kịp thời - Hỗ trợ lớp trực tuyến nếu gặp vấn đề	- Nhật ký trực ca tối - Chăm sóc học viên - Hệ thống đào tạo nội bộ - Thông kê lỗi lớp trực tuyến	- Check zalo trước để hỗ trợ cho những học viên học ca sáng - Xử lý vấn đề tồn đọng của ca tối - Trao đổi với bộ phận kỹ thuật về các lỗi chưa khắc phục được
10h00 - 11h00	- Tiếp nhận học viên mới - Check Zalo để hỗ trợ học viên kịp thời - Hỗ trợ lớp trực tuyến nếu gặp vấn đề	- Chăm sóc học viên - Hệ thống đào tạo nội bộ	- Sau khi có thông tin của học viên thì gửi mail "hướng dẫn kiểm tra thiết bị" và kết bạn zalo - Xin lịch học của học viên và trao đổi với giáo viên để mở lớp
11h00 - 12h00	- Check buổi cho học viên - Check Zalo để hỗ trợ học viên kịp thời - Hỗ trợ lớp trực tuyến nếu gặp vấn đề	- Chăm sóc học viên - Hệ thống đào tạo nội bộ	- Kiểm tra lịch sử giảng dạy và zalo để nắm rõ số buổi đã học của các lớp
13h30 - 15h00	- Tiến hành mở lớp cho học viên - Check Zalo để hỗ trợ học viên kịp thời - Hỗ trợ lớp trực tuyến nếu gặp vấn đề	- Chăm sóc học viên - Hệ thống đào tạo nội bộ - File lịch dạy của giáo viên	- Sau khi có đầy đủ thông tin thì tạo lớp cho học viên (Lưu ý thêm đầy đủ vào file giáo viên và hoàn thành đủ các thao tác)
15h00 - 16h00	- Gửi thông tin cho học viên và giáo viên - Check Zalo để hỗ trợ học viên kịp thời - Hỗ trợ lớp trực tuyến nếu gặp vấn đề	- Zalo, line	- Gửi đầy đủ thông tin cho giáo viên và học viên - Hướng dẫn học viên đăng nhập vào hệ thống (Nếu cần)
16h00 - 17h00	- Giải quyết những vấn đề tồn đọng trong buổi sáng - Check Zalo để hỗ trợ học viên kịp thời - Hỗ trợ lớp trực tuyến nếu gặp vấn đề	- Nhật ký trực ca tối - Chăm sóc học viên - Hệ thống đào tạo nội bộ - Thông kê lỗi lớp trực tuyến	- Cố gắng giải quyết gọn những vấn đề của file "nhật ký" và file "thống kê lỗi" để tránh buổi học tới có vấn đề
17h00 - 18h00	- Note thông tin vào file "Nhật ký trực ca tối" - Check Zalo để hỗ trợ học viên kịp thời - Hỗ trợ lớp trực tuyến nếu gặp vấn đề	- Nhật ký trực ca tối - Chăm sóc học viên - Hệ thống đào tạo nội bộ - Thông kê lỗi lớp trực tuyến	- Note thông tin "lớp học buổi đầu" vào file "nhật ký" để các bạn check và hỗ trợ học viên

Hình 2.6 Quy trình chăm sóc học viên (Nguồn: Công ty Edura)

2.3.2. Quy trình quản lý bán hàng

Phòng Sale sẽ tiếp nhận thông tin từ phòng Marketing, sau đó tiến hành tư vấn khóa học phù hợp với khách hàng.

Để có thể sắp xếp được khóa học phù hợp với trình độ của khách hàng, nhân viên sale sẽ tiến hành đặt lịch test với giáo viên chuyên môn.

Sau khi đã biết được trình độ của học viên, nhân viên sale sẽ tiến hành xác định lộ trình học tập, giá lợ trình và thông báo đến khách hàng.

Khách hàng đồng ý với lộ trình đã được tư vấn sẽ tiến hành đăng ký học để trở thành học viên của trung tâm.

Sau khi hoàn thành hồ sơ thủ tục đăng ký học viên sẽ tiến học thanh toán học phí và bàn giao sang cho phòng đào tạo để tiếp tục quy trình quản lý học của học viên.

CHƯƠNG 3. GIỚI THIỆU VÀ XÂY DỰNG CÁC HỆ THỐNG NGUỒN DỮ LIỆU

Hệ thống gồm 3 nguồn thuộc sự quản lý của 3 phòng ban khác nhau: Phòng quản lý (Phòng đào tạo), Phòng Sale, Phòng Kế toán.

3.1. Hệ thống nguồn 1 (SQL Server - Phòng đào tạo)

Hệ thống nguồn 1 là Database được xây dựng trên Microsoft SQL Server, quản lý bởi phòng đào tạo. Dữ liệu được thiết kế như sau:

3.1.1. Thiết kế cơ sở dữ liệu mức khái niệm

3.1.1.1. Xác định thực thể

- Học viên.
- Khoa học.
- Giáo viên.
- Nhân viên Sale.

3.1.1.1.1. Xác định mối quan hệ thực thể

- HOCHIEP (ID_HV, HoTen, GioiTinh, Sdt, Email, NgaySinh, NgheNghiep).
- KHOAHOC (ID_KH, TenKH, hoc phò, GiaKM, SoBuoiHoc, LoaiKH).
- GIAOVIEN (ID_GV, HoTen, GioiTinh, Sdt, QuocTich, MucLuong).
- NVSALE (ID_NV, HoTen, GioiTinh).

3.1.2. Thiết kế cơ sở dữ liệu mức logic

3.1.2.1. Chuyển hóa quan hệ thành các quan hệ

Quan hệ DANGKY, THANHTOAN, HOC là các quan hệ bậc 3 có thuộc tính riêng nên được chuyển thành quan hệ mới:

- DANGKY (ID_NV, ID_HV, ID_KH, NgayDK).
- THANHTOAN (ID_HV, ID_KH, TrangthaiTT, NgayTT, NH_Nop, NH_Nhan - phòng sale, SoTien).

- HOC (ID, ID_KH, ID_HV, ID_GV, TrangthaiLophoc, NgayBD, NgayKT, SobuoiConlai, GioHoc, NgayHoc).

3.1.2.2. Chuẩn hóa thực thể

- HOCHIEN (ID_HV, HoTen, GioiTinh, Sdt, Email, NgaySinh, NgheNghiep).
- KHOAHOC (ID_KH, TenKH, GiaKM, SoBuoiHoc, LoaiKH).
- GIAOVIEN (ID_GV, HoTen, GioiTinh, Sdt, QuocTich, MucLuong).
- NVSALE (ID_NV, HoTen, GioiTinh).
- DANGKY (ID_NV, ID_HV, ID_KH, NgayDK).
- THANHTOAN (ID_HV, ID_KH, TrangthaiTT, NgayTT, NH_Nop, NH_Nhan, SoTien).
- HOC (ID, ID_KH, ID_HV, ID_GV, TrangthaiLophoc, NgayBD, NgayKT, SobuoiConlai, GioHoc, NgayHoc).

3.1.3. Thiết kế cơ sở dữ liệu mức vật lý

HOCHIEN					
Tên cột	Cột dữ liệu	Kích thước	Định dạng	Allows Null	Ràng buộc
ID_HV	varchar	20		NOT NULL	Khóa chính
HoTen	nvarchar	100		NOT NULL	
GioiTinh	nvarchar	5			
NgaySinh	date	10			
Sdt	varchar	10			
Email	varchar	100			
NgheNghiep	nvarchar	100			

Bảng 3.1 Cơ sở dữ liệu mức vật lý bảng Học Viên

GIAOVIEN					
Tên cột	Cột dữ liệu	Kích thước	Định dạng	Allows Null	Ràng buộc
ID_GV	varchar	20		NOT NULL	Khóa chính
HoTen	nvarchar	100		NOT NULL	
GioiTinh	nvarchar	5			
QuocTich	nvarchar	20			
Sdt	varchar	10			
MucLuong	int				

Bảng 3.2 Cơ sở dữ liệu mức vật lý bảng Giáo Viên

KHOAHOC					
Tên cột	Cột dữ liệu	Kích thước	Định dạng	Allows Null	Ràng buộc
ID_KH	varchar	20		NOT NULL	Khóa chính
TenKH	nvarchar	100		NOT NULL	
GiaKM	float				
SoBuoiHoc	int				
LoaiKH	nvarchar	10			

Bảng 3.3 Cơ sở dữ liệu mức vật lý bảng Khóa Học

NVSALE					
Tên cột	Cột dữ liệu	Kích thước	Định dạng	Allows Null	Ràng buộc
ID_NV	varchar	20		NOT NULL	Khóa chính
HoTen	nvarchar	100		NOT NULL	
GioiTinh	nvarchar	5			

Bảng 3.4 Cơ sở dữ liệu mức vật lý bảng Nhân Viên Sale

DANGKY					
Tên cột	Cột dữ liệu	Kích thước	Định dạng	Allows Null	Ràng buộc
ID_NV	varchar	20		NOT NULL	Khóa chính
ID_HV	varchar	20		NOT NULL	Khóa chính
ID_KH	varchar	20		NOT NULL	Khóa chính
NgayDK	date				

Bảng 3.5 Cơ sở dữ liệu mức vật lý bảng Đăng ký

HOC					
Tên cột	Cột dữ liệu	Kích thước	Định dạng	Allows Null	Ràng buộc
ID	varchar	20		NOT NULL	Khóa chính
ID_HV	varchar	20		NOT NULL	Khóa chính
ID_KH	varchar	20		NOT NULL	Khóa chính

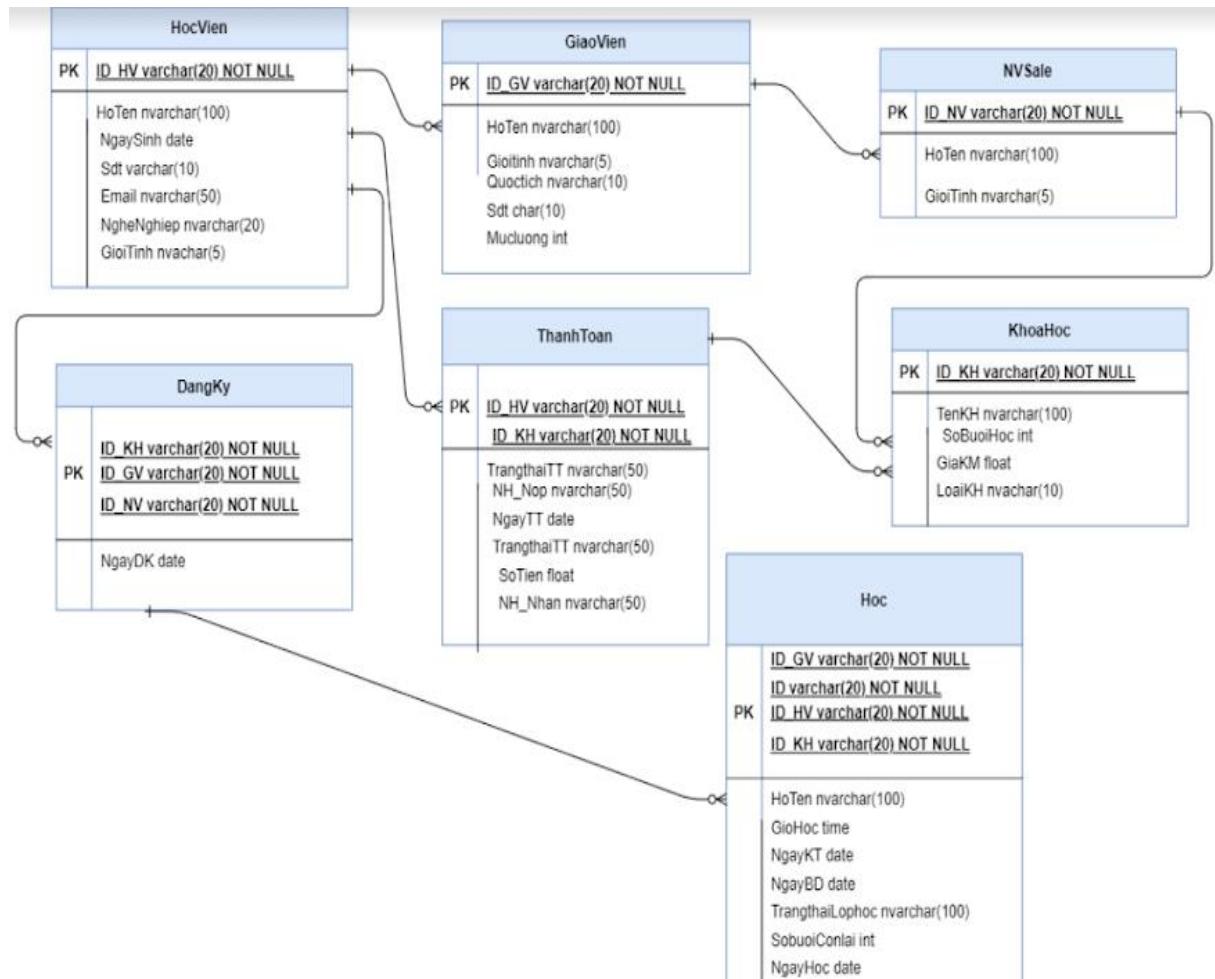
ID_GV	varchar	20		NOT NULL	Khóa chính
NgayBD	date			NOT NULL	
NgayKT	date			NOT NULL	
TrangthaiLophoc	nvarchar	100		NOT NULL	
SobuoiConlai	int			NOT NULL	
GioHoc	time			NOT NULL	
NgayHoc	date			NOT NULL	

Bảng 3.6 Cơ sở dữ liệu mức vật lý bảng Học

THANHTOAN					
Tên cột	Cột dữ liệu	Kích thước	Định dạng	Allows Null	Ràng buộc
ID_HV	varchar	20		NOT NULL	Khóa chính
ID_KH	varchar	20		NOT NULL	Khóa chính
TrangthaiTT	nvarchar	50		NOT NULL	
NgayTT	date			NOT NULL	
NH_Nop	nvarchar	50		NOT NULL	
NH_Nhan	nvarchar	50		NOT NULL	
SoTien	float			NOT NULL	

Bảng 3.7 Cơ sở dữ liệu mức vật lý bảng Thanh Toán

3.1.4. Mô hình dữ liệu quan hệ đã chuẩn hóa



Hình 3.1 Mô hình dữ liệu quan hệ đã chuẩn hóa

3.1.5. Trigger tự động tính số buổi còn lại

```
create trigger Tinh_So_Buoi_Con_Lai
on hoc
after insert as
BEGIN
    UPDATE hoc
    SET hoc.sobuoiconlai = hoc.tongbobuoi - hoc.sobuoidahoc
    FROM hoc
END
```

3.1.6. Dữ liệu nguồn 1

Nguồn 1 thông tin học viên được lưu dưới dạng Họ và tên, Giới tính: Nam/Nữ.

ID_HV	Hoten	Gioitinh	Sdt	Email	TrinhDo
183	112019101	Trần Thị Thanh Hiền	Nữ	905476274	thanhienlove81.dng@gmail.com
184	112019102	Huỳnh Thị Âu	Nữ	812771106	auhuynh1010@gmail.com
185	112019103	Nguyễn Thị Bảo Trân	Nữ	915227768	trannguyen0287@gmail.com
186	112019104	nguyễn hồng nhung	Nữ	7075618152	honghungnhung801@gmail.com
187	112019105	Nguyễn Thị Ngoan	Nữ	356102913	ngoannguyenkt.1990@gmail
188	112019106	Nguyễn Thị Giang	Nữ	353044358	giangnguyen9h4@gmail.com
189	112019107	phan thanh sang	Nam	368355760	phanthansang1093@gmail.com
190	112019108	Nguyễn Đào Trung	Nam	918396226	nguyendaotrung0408@gmail.com
191	112019109	mai thị thường	Nữ	384085130	maithuongjpp1991@gmail.com
192	112019110	Phạm Hà Linh	Nữ	962501874	phamhalinh12@gmail.com
193	112019111	Lã Thị Trang	Nữ	888463518	latrangtenshin@gmail.com
194	112019112	Nguyễn Thị Hồng	Nữ	982331035	hongtra1992@gmail.com
195	112019113	Phan Văn Quyết	Nam	978902291	phanquyetx5@gmail.com
196	112019116	Đỗ Thanh Thuỷ	Nữ	357942424	dothiloi24041998@gmail.com
197	112019117	Nguyễn Quốc	Nam	394402950	nguyenquoc.tink32c@gmail.com
198	112019118	Lê Văn Anh	Nữ	948397734	levananh2994@gmail.com
199	112019119	Nguyễn Thanh Thảo	Nam	369138770	thaont16@fsoft.com.vn
200	112019120	Thái Thu Hà	Nữ	984220984	thaithuha@gmail.com
201	112019121	Nguyễn Thủ	Nữ	973530225	anhtho1602@gmail.com
202	112019122	Nguyễn Duy Hào	Nam	349964097	trangvondh@gmail.com

Hình 3.2 Nguồn dữ liệu 1

3.2. Hệ thống nguồn 2 (Excel File - Phòng Sale)

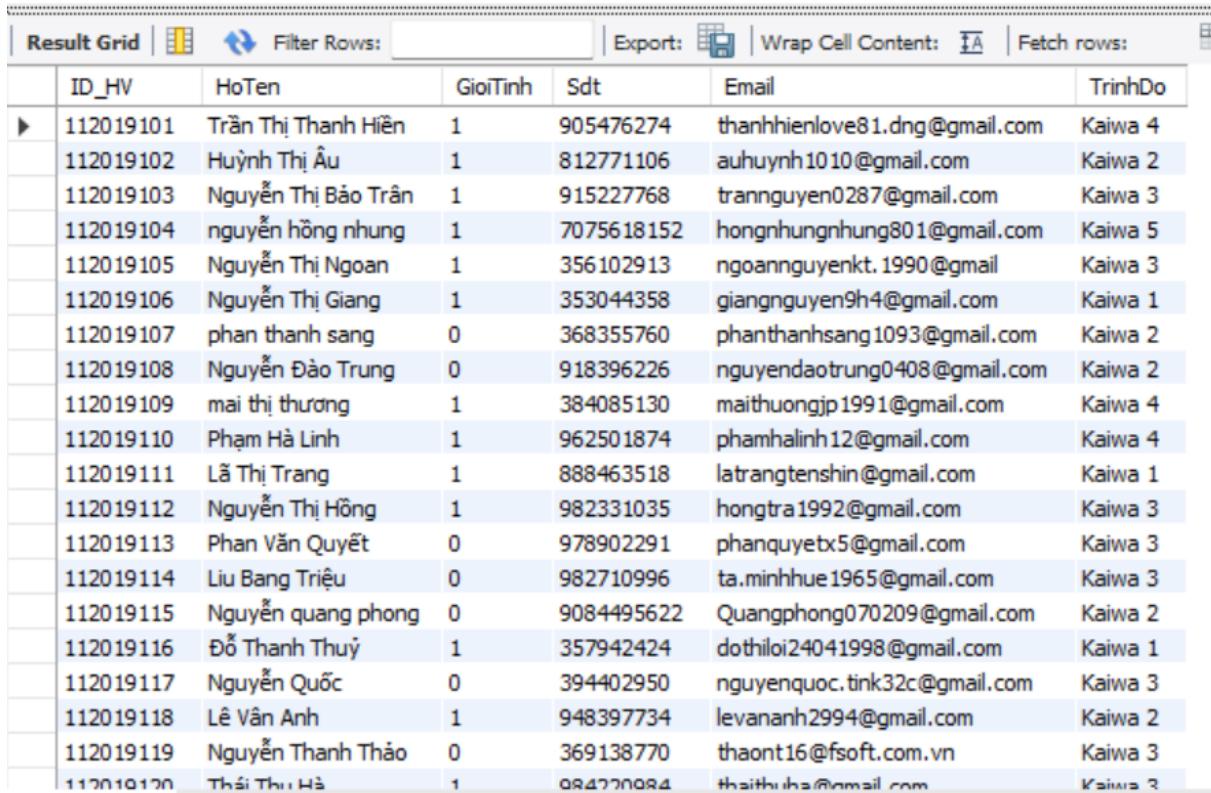
Hệ thống Nguồn 2 là file Excel thuộc phòng Sale chứa thông tin của học viên tham gia học tại trung tâm như id học viên, họ đệm, tên, giới tính, số điện thoại, email, trình độ.

ID_HV	Họ đệm	Tên	Giới tính	Sđt	Email	Trình độ
72019001	Mi Xa	Tran	Female	917520969	mixa.brvt@gmail.com	Kaiwa 1
72019002	Hoang Thi	Phuong	Female	842323225	phuongnguyena@gmail.com	Kaiwa 3
72019003	Trần Việt	Hạnh	Female	914297679	viethanh.dn@gmail.com	Kaiwa 1
72019004	Nguyễn Thái	Hòa	Female	906220311	thaihoafstu@gmail.com	Kaiwa 3
72019005	Duy	Thành	Male	977701689	Duythanhdcn4@gmail.com	Kaiwa 5
72019006	Duy	Bùi	Male	911263017	Builetuduy@gmail.com	Kaiwa 2
72019007	Trần Vinh	Hội	Male	386330057	vinhhoi2005@gmail.com	Kaiwa 4
72019008	Phạm Khánh	Ngọc	Female	914869998	khanhngoc0208@gmail.com	Kaiwa 5
72019009	Trần minh	trung	Male	869690927	tmtrung2010199@gmail.com	Kaiwa 1
72019010	Đức	Mạnh	Male	356442269	ng.manh.8888@gmail.com	Kaiwa 3
72019011	Vũ Thị Hằng	Nga	Female	978436072	sibobaka@gmail.com	Kaiwa 3
72019012	Lê Ngọc	Trinh	Female	764107695	lengoctrinh0312@gmail.com	Kaiwa 2
72019013	Vũ Văn	Anh	Male	368614246	vuvanken111894@gmail.com	Kaiwa 3
72019014	Võ Thị	Trà	Female	989539440	tra.vo2706@gmail.com	Kaiwa 1
72019015	Vũ Đinh	Phong	Male	8041811792	haphong210520@gmail.com	Kaiwa 2
72019016	Phí Văn	Hoàng	Male	903465700	apexhoang@gmail.com	Kaiwa 1
72019017	Nguyễn Thị Tô	Lịch	Female	985306693	lequelinhchi168@gmail.com	Kaiwa 1
72019018	Nguyễn thị	thuỷ	Female	1699932708	phamquynhanh0@gmail.com	Kaiwa 1
72019019	Nguyễn Huỳnh Phương	Thảo	Female	903049822	nhpt017@gmail.com	Kaiwa 1

Hình 3.3 Dữ liệu trên Excel

3.3. Hệ thống nguồn 3 (MySQL Source)

Hệ thống nguồn 3 (MySQL Source) xây dựng trên MySQL được quản lý bởi bộ phận kế toán bao gồm 2 bảng là bảng học viên. Khác với hệ thống 1 và hệ thống 2 tại đây dữ liệu về giới tính được lưu dưới dạng 0/1 (0 - Nam, 1 - Nữ).



The screenshot shows a MySQL Workbench interface with a result grid titled 'Result Grid'. The grid displays 20 rows of student information with the following columns: ID_HV, HoTen, GioiTinh, Sdt, Email, and TrinhDo. The data includes various student names, gender (1 for male, 0 for female), phone numbers, emails, and their respective levels (Kaiwa 1 through Kaiwa 5). The 'Email' column contains several Gmail addresses.

	ID_HV	HoTen	GioiTinh	Sdt	Email	TrinhDo
▶	112019101	Trần Thị Thanh Hiền	1	905476274	thanhhiendng@gmail.com	Kaiwa 4
	112019102	Huỳnh Thị Âu	1	812771106	ahuynh1010@gmail.com	Kaiwa 2
	112019103	Nguyễn Thị Bảo Trân	1	915227768	trannguyen0287@gmail.com	Kaiwa 3
	112019104	nguyễn hồng nhung	1	7075618152	honghungnhung801@gmail.com	Kaiwa 5
	112019105	Nguyễn Thị Ngoan	1	356102913	ngoannguyenkt.1990@gmail	Kaiwa 3
	112019106	Nguyễn Thị Giang	1	353044358	giangnguyen9h4@gmail.com	Kaiwa 1
	112019107	phan thanh sang	0	368355760	phanthangsang1093@gmail.com	Kaiwa 2
	112019108	Nguyễn Đào Trung	0	918396226	nguyendaotrung0408@gmail.com	Kaiwa 2
	112019109	mai thị thương	1	384085130	maithuongjp1991@gmail.com	Kaiwa 4
	112019110	Phạm Hà Linh	1	962501874	phamhalinh12@gmail.com	Kaiwa 4
	112019111	Lã Thị Trang	1	888463518	latrangtenshin@gmail.com	Kaiwa 1
	112019112	Nguyễn Thị Hồng	1	982331035	hongtra1992@gmail.com	Kaiwa 3
	112019113	Phan Văn Quyết	0	978902291	phanquyetx5@gmail.com	Kaiwa 3
	112019114	Liu Bang Triệu	0	982710996	ta.minhhue1965@gmail.com	Kaiwa 3
	112019115	Nguyễn quang phong	0	9084495622	Quangphong070209@gmail.com	Kaiwa 2
	112019116	Đỗ Thanh Thuỷ	1	357942424	dothiloi24041998@gmail.com	Kaiwa 1
	112019117	Nguyễn Quốc	0	394402950	nguyenquoc.tink32c@gmail.com	Kaiwa 3
	112019118	Lê Văn Anh	1	948397734	levanh2994@gmail.com	Kaiwa 2
	112019119	Nguyễn Thanh Thảo	0	369138770	thaont16@fsoft.com.vn	Kaiwa 3
	112019120	Thái Thị Hà	1	094770084	thaititha@gmail.com	Kaiwa 3

Hình 3.4 Hệ thống nguồn thứ 3 (MySQL Source)

Tại hệ thống nguồn 3: học viên Trần Thị Thanh Hiền với ID = 112019101 vẫn còn ở trình độ Kaiwa 4 khác với Kaiwa 5 của phòng đào tạo (do học viên học tại Trung tâm nên đã được giúp nâng trình độ lên Kaiwa 5 nhưng dữ liệu của phòng kế toán không được cập nhật).

CHƯƠNG 4. THIẾT KẾ VÀ MÔ TẢ CÁC BẢNG DIM, FACT

4.1. Nghệp vụ chính gắn với các bảng Dim, Fact trong Data Mart

Data Mart là phiên bản đơn giản hóa của Data warehouse - cung cấp cho người dùng dữ liệu cụ thể về một bộ phận của tổ chức hoặc một doanh nghiệp. Chức năng chính của data mart là thường đưa ra những thông tin liên quan cần thiết để đưa ra quyết định quan trọng trong những bộ phận cụ thể của doanh nghiệp.

- ⇒ Nhóm đã chọn và phân tích data mart gồm các bảng dim, fact liên quan đến 2 quy trình nghiệp vụ là quản lý học tập và quản lý đăng ký học và thanh toán của học viên (Cách tiếp cận Bottom-up).
- Data mart 1 - quy trình quản lý học tập của học viên bao gồm: Fact_Hoc, Dim_GiaoVien, Dim_KhoaHoc, Dim_HocVien, Dim_Date.
- Data mart 2 - quy trình quản lý đăng ký học - thanh toán của học viên gồm: Fact_Sales, Dim_NVSale, Dim_KhoaHoc, Dim_Thanhtoan_bị, Dim_Dangky, Dim_HocVien, Dim_Date.

4.2. Lựa chọn kiểu thiết kế phù hợp cho bảng Fact

Vì bảng fact mô tả sự kiện xảy ra trong thế giới thực và được ghi vào Data Warehouse. Các tiêu chí và trường dữ liệu trong bảng này không mô tả quá trình, ghi nhận thời điểm xảy ra sự kiện. Bảng này có quy mô số lượng và chi tiết lớn nhất. Đây là đầu vào để tổng hợp lên các bảng fact có mức tổng hợp cao hơn.

- ⇒ Vì vậy, nhóm lựa chọn thiết kế bảng Fact theo kiểu giao dịch.

4.3. Mô tả rõ ý nghĩa các trường dữ liệu và nguồn của mỗi bảng Dim, Fact

4.3.1. Xây dựng bảng cắt lớp thời gian Dim_Date

Thời gian là một chiều rất quan trọng được dùng trong hầu hết mọi bảng fact, dim_date thường được tổ chức đặc biệt và không có nguồn nhập. Dim thời gian thường được dùng dạng tham chiếu cho nhiều chiều khác nhau. Trong dự án, nhóm xác định

Dim_Date là bảng không thể thiếu trong kho dữ liệu hỗ trợ cho việc quản lý học tập của học viên.

Các lựa chọn về thời gian như thứ, ngày, tháng quý, ngày nghỉ,... giúp dễ dàng quản lý lịch dạy của trung tâm, tạo các báo cáo theo chu kỳ. Đặc biệt Holiday giúp công ty phân biệt ngày thường ngày lễ dễ dàng đưa ra các chương trình khuyến mãi khóa học (ví dụ như Ngày nhà giáo Việt Nam, ngày Tết dương,...).

Dim_Date	
!	Date_id
	Date
	Day
	DaySuffix
	Weekday
	WeekDayName
	WeekDayName_Short
	WeekDayName_FirstLetter
	DOWInMonth
	DayOfYear
	WeekOfMonth
	WeekOfYear
	Month
	MonthName
	MonthName_Short
	MonthName_FirstLetter
	Quarter
	QuarterName
	Year
	MMYYYY
	MonthYear
	IsWeekend
	IsHoliday
	HolidayName
	FirstDateofYear
	LastDateofYear
	FirstDateofQuater
	LastDateofQuater
	FirstDateofMonth
	LastDateofMonth
	FirstDateofWeek
	LastDateofWeek

Hình 4.1 Bảng Dim_Date

Bảng Dim_date sẽ được tự động bằng câu lệnh SQL với thời gian từ 01/01/2018 và kết thúc 31/12/2030. Câu lệnh tạo bảng Dim_Date:

```
--tao bang Dim_Date
CREATE TABLE [dbo].[Dim_Date] (
    [Date_id] [int] NOT NULL,
    [Date] [date] NOT NULL,
    [Day] [int] NOT NULL,
    [DaySuffix] [char](2) NOT NULL, --Áp dụng hậu tố như 1st, 2nd, 3rd, ...
    [Weekday] [tinyint] NOT NULL,
    [WeekDayName] [varchar](10) NOT NULL,
    [WeekDayName_Short] [char](3) NOT NULL,
    [WeekDayName_FirstLetter] [char](1) NOT NULL,
    [DOWInMonth] [tinyint] NOT NULL, --Số ngày trong tháng
    [DayOfYear] [smallint] NOT NULL,
    [WeekOfMonth] [tinyint] NOT NULL,
    [WeekOfYear] [tinyint] NOT NULL,
    [Month] [tinyint] NOT NULL,
    [MonthName] [varchar](10) NOT NULL,
    [MonthName_Short] [char](3) NOT NULL,
    [MonthName_FirstLetter] [char](1) NOT NULL,
    [Quarter] [tinyint] NOT NULL,
    [QuarterName] [varchar](6) NOT NULL, --First, second, ...
    [Year] [int] NOT NULL,
    [MMYYYY] [char](6) NOT NULL,
    [MonthYear] [char](7) NOT NULL, --Jan-2022
    IsWeekend BIT NOT NULL, --0- ngày thường, 1- T7/CN
    IsHoliday BIT NOT NULL, --0: Ngày thường, 1- ngày lễ
    HolidayName NVARCHAR(50) NULL,
    [FirstDateofYear] DATE NULL,
    [LastDateofYear] DATE NULL,
    [FirstDateofQuater] DATE NULL,
    [LastDateofQuater] DATE NULL,
    [FirstDateofMonth] DATE NULL,
    [LastDateofMonth] DATE NULL,
    [FirstDateofWeek] DATE NULL,
    [LastDateofWeek] DATE NULL,
    PRIMARY KEY CLUSTERED ([Date_id] ASC));

--inset into Dim_Date
SET NOCOUNT ON

DECLARE @StartDate DATE = '2018-01-01'
DECLARE @EndDate DATE = '2030-12-31'
WHILE @StartDate < @EndDate
BEGIN
    INSERT INTO [dbo].[Dim_Date] (
        [Date_id],
        [Date],
        [Day],
```

```

[DaySuffix],
[Weekday],
[WeekDayName],
[WeekDayName_Short],
[WeekDayName_FirstLetter],
[DOWInMonth],
[DayOfYear],
[WeekOfMonth],
[WeekOfYear],
[Month],
[MonthName],
[MonthName_Short],
[MonthName_FirstLetter],
[Quarter],
[QuarterName],
[Year],
[MMYYYY],
[MonthYear],
[IsWeekend],
[IsHoliday],
[FirstDateofYear],
[LastDateofYear],
[FirstDateofQuater],
[LastDateofQuater],
[FirstDateofMonth],
[LastDateofMonth],
[FirstDateofWeek],
[LastDateofWeek]
)
SELECT Date_id = YEAR(@StartDate) * 10000 + MONTH(@StartDate) * 100 +
DAY(@StartDate),
DATE = @StartDate,
Day = DAY(@StartDate),
[DaySuffix] = CASE
WHEN DAY(@StartDate) = 1
    OR DAY(@StartDate) = 21
    OR DAY(@StartDate) = 31
    THEN 'st'
WHEN DAY(@StartDate) = 2
    OR DAY(@StartDate) = 22
    THEN 'nd'
WHEN DAY(@StartDate) = 3
    OR DAY(@StartDate) = 23
    THEN 'rd'
ELSE 'th'
END,
WEEKDAY = DATEPART(dw, @StartDate),
WeekDayName = DATENAME(dw, @StartDate),
WeekDayName_Short = UPPER(LEFT(DATENAME(dw, @StartDate), 3)),
WeekDayName_FirstLetter = LEFT(DATENAME(dw, @StartDate), 1),

```

```

[DOWInMonth] = DAY(@StartDate),
[DayOfYear] = DATENAME(dy, @StartDate),
[WeekOfMonth] = DATEPART(WEEK, @StartDate) - DATEPART(WEEK,
DATEADD(MM, DATEDIFF(MM, 0, @StartDate), 0)) + 1,
[WeekOfYear] = DATEPART(wk, @StartDate),
[Month] = MONTH(@StartDate),
[MonthName] = DATENAME(mm, @StartDate),
[MonthName_Short] = UPPER(LEFT(DATENAME(mm, @StartDate), 3)),
[MonthName_FirstLetter] = LEFT(DATENAME(mm, @StartDate), 1),
[Quarter] = DATEPART(q, @StartDate),
[QuarterName] = CASE
    WHEN DATENAME(qq, @StartDate) = 1
        THEN 'first'
    WHEN DATENAME(qq, @StartDate) = 2
        THEN 'second'
    WHEN DATENAME(qq, @StartDate) = 3
        THEN 'third'
    WHEN DATENAME(qq, @StartDate) = 4
        THEN 'fourth'
    END,
[Year] = YEAR(@StartDate),
[MMYYYY] = RIGHT('0' + CAST(MONTH(@StartDate) AS VARCHAR(2)), 2) +
CAST(YEAR(@StartDate) AS VARCHAR(4)),
[MonthYear] = CAST(YEAR(@StartDate) AS VARCHAR(4)) +
UPPER(LEFT(DATENAME(mm, @StartDate), 3)),
[IsWeekend] = CASE
    WHEN DATENAME(dw, @StartDate) = 'Sunday'
        OR DATENAME(dw, @StartDate) = 'Saturday'
        THEN 1
    ELSE 0
    END,
[IsHoliday] = 0,
--[IsSpecialDay] = 0,
[FirstDateofYear] = CAST(CAST(YEAR(@StartDate) AS VARCHAR(4)) + '-01-
01' AS DATE),
[LastDateofYear] = CAST(CAST(YEAR(@StartDate) AS VARCHAR(4)) + '-12-
31' AS DATE),
[FirstDateofQuater] = DATEADD(qq, DATEDIFF(qq, 0, GETDATE()), 0),
[LastDateofQuater] = DATEADD(dd, - 1, DATEADD(qq, DATEDIFF(qq, 0,
GETDATE()) + 1, 0)),
[FirstDateofMonth] = CAST(CAST(YEAR(@StartDate) AS VARCHAR(4)) + '-' +
CAST(MONTH(@StartDate) AS VARCHAR(2)) + '-01' AS DATE),
[LastDateofMonth] = CONVERT(DATETIME, CONVERT(DATE, DATEADD(DD, -
(DATEPART(DD, (DATEADD(MM, 1, @StartDate)))), DATEADD(MM, 1, @StartDate))), -
--[LastDateofMonth] = EOMONTH(@StartDate),
[FirstDateofWeek] = DATEADD(dd, - (DATEPART(dw, @StartDate) - 1),
@StartDate),
[LastDateofWeek] = DATEADD(dd, 7 - (DATEPART(dw, @StartDate)),
@StartDate)

```

```

    SET @StartDate = DATEADD(DD, 1, @StartDate)
END

UPDATE Dim_Date
SET [IsHoliday] = 1,
[HolidayName] = N'Tết dương lịch'
WHERE [Month] = 1
AND [DAY] = 1;

UPDATE Dim_Date
SET [IsHoliday] = 1,
[HolidayName] = N'30 Tết Âm lịch 2018'
WHERE [YEAR] = 2018
AND [Month] = 2
AND [DAY] = 15;

UPDATE Dim_Date
SET [IsHoliday] = 1,
[HolidayName] = N'Mùng 1 Tết Âm lịch 2018'
WHERE [YEAR] = 2018
AND [Month] = 2
AND [DAY] = 16;

UPDATE Dim_Date
SET [IsHoliday] = 1,
[HolidayName] = N'Giỗ Tổ Hùng Vương năm 2018'
WHERE [YEAR] = 2018
AND [Month] = 4
AND [DAY] = 25;

UPDATE Dim_Date
SET [IsHoliday] = 1,
[HolidayName] = N'30 Tết Âm lịch 2019'
WHERE [YEAR] = 2019
AND [Month] = 2
AND [DAY] = 4;

UPDATE Dim_Date
SET [IsHoliday] = 1,
[HolidayName] = N'Mùng 1 Tết Âm lịch 2019'
WHERE [YEAR] = 2019
AND [Month] = 2
AND [DAY] = 5;

UPDATE Dim_Date
SET [IsHoliday] = 1,
[HolidayName] = N'Giỗ Tổ Hùng Vương năm 2019'
WHERE [YEAR] = 2019
AND [Month] = 4
AND [DAY] = 14;

UPDATE Dim_Date
SET [IsHoliday] = 1,
[HolidayName] = N'30 Tết Âm lịch 2020'

```

```

WHERE [YEAR] = 2020
    AND [Month] = 1
    AND [DAY] = 24;
UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Mùng 1 Tết Âm lịch 2020'
WHERE [YEAR] = 2020
    AND [Month] = 1
    AND [DAY] = 25;
UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Giỗ Tổ Hùng Vương năm 2020'
WHERE [YEAR] = 2020
    AND [Month] = 4
    AND [DAY] = 2;

UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'30 Tết Âm lịch 2021'
WHERE [YEAR] = 2021
    AND [Month] = 2
    AND [DAY] = 11;
UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Mùng 1 Tết Âm lịch 2021'
WHERE [YEAR] = 2021
    AND [Month] = 2
    AND [DAY] = 12;
UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Giỗ Tổ Hùng Vương năm 2021'
WHERE [YEAR] = 2021
    AND [Month] = 4
    AND [DAY] = 21;

UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'30 Tết Âm lịch 2022'
WHERE [YEAR] = 2022
    AND [Month] = 1
    AND [DAY] = 31;
UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Mùng 1 Tết Âm lịch 2022'
WHERE [YEAR] = 2022
    AND [Month] = 2
    AND [DAY] = 1;
UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Giỗ Tổ Hùng Vương năm 2022'

```

```

WHERE [YEAR] = 2022
AND [Month] = 4
AND [DAY] = 8;

UPDATE Dim_Date
SET [IsHoliday] = 1,
[HolidayName] = N'Valentine 14/2'
WHERE [Month] = 2
AND [DAY] = 14;

UPDATE Dim_Date
SET [IsHoliday] = 1,
[HolidayName] = N'Ngày quốc tế phụ nữ 8/3'
WHERE [Month] = 3
AND [DAY] = 8;

UPDATE Dim_Date
SET [IsHoliday] = 1,
[HolidayName] = N'Ngày Giải phóng Miền Nam 30/4'
WHERE [Month] = 4
AND [DAY] = 30;

UPDATE Dim_Date
SET [IsHoliday] = 1,
[HolidayName] = N'Ngày Quốc tế Lao động 1/5'
WHERE [Month] = 5
AND [DAY] = 1;

UPDATE Dim_Date
SET [IsHoliday] = 1,
[HolidayName] = N'Ngày Quốc tế thiếu nhi 1/6'
WHERE [Month] = 6
AND [DAY] = 1;

UPDATE Dim_Date
SET [IsHoliday] = 1,
[HolidayName] = N'Ngày Quốc khánh Việt Nam 2/9'
WHERE [Month] = 9
AND [DAY] = 2;

UPDATE Dim_Date
SET [IsHoliday] = 1,
[HolidayName] = N'Ngày phụ nữ Việt Nam 20/10'
WHERE [Month] = 10
AND [DAY] = 10;

UPDATE Dim_Date
SET [IsHoliday] = 1,
[HolidayName] = N'Ngày thành lập công ty Edura'

```

```

WHERE [Month] = 8
    AND [DAY] = 9;

UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Ngày Nhà giáo Việt Nam 20/11'
WHERE [Month] = 11
    AND [DAY] = 20;

UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Giáng sinh'
WHERE [Month] = 12
    AND [DAY] = 25;

UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'30 Tết Âm lịch 2023'
WHERE [YEAR] = 2023
    AND [Month] = 1
    AND [DAY] = 21;
UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Mùng 1 Tết Âm lịch 2023'
WHERE [YEAR] = 2023
    AND [Month] = 22
    AND [DAY] = 1;

UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Giỗ Tổ Hùng Vương 2023'
WHERE [YEAR] = 2023
    AND [Month] = 4
    AND [DAY] = 29;

UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'30 Tết Âm lịch 2024'
WHERE [YEAR] = 2024
    AND [Month] = 2
    AND [DAY] = 9;
UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Mùng 1 Tết Âm lịch 2024'
WHERE [YEAR] = 2024
    AND [Month] = 2
    AND [DAY] = 10;

UPDATE Dim_Date
SET [IsHoliday] = 1,

```

```

[HolidayName] = N'Giỗ Tổ Hùng Vương 2024'
WHERE [YEAR] = 2024
    AND [Month] = 4
    AND [DAY] = 18;

UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'30 Tết Âm lịch 2025'
WHERE [YEAR] = 2025
    AND [Month] = 1
    AND [DAY] = 28;
UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Mùng 1 Tết Âm lịch 2025'
WHERE [YEAR] = 2025
    AND [Month] = 1
    AND [DAY] = 29;

UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Giỗ Tổ Hùng Vương 2025'
WHERE [YEAR] = 2025
    AND [Month] = 4
    AND [DAY] = 7;

UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'30 Tết Âm lịch 2026'
WHERE [YEAR] = 2026
    AND [Month] = 2
    AND [DAY] = 16;
UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Mùng 1 Tết Âm lịch 2026'
WHERE [YEAR] = 2026
    AND [Month] = 2
    AND [DAY] = 17;

UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Giỗ Tổ Hùng Vương 2026'
WHERE [YEAR] = 2026
    AND [Month] = 4
    AND [DAY] = 26;

UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'29 Tết Âm lịch 2027'
WHERE [YEAR] = 2027
    AND [Month] = 2

```

```

        AND [DAY] = 5;
UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Mùng 1 Tết Âm lịch 2027'
WHERE [YEAR] = 2027
    AND [Month] = 2
    AND [DAY] = 6;

UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Giỗ Tổ Hùng Vương 2027'
WHERE [YEAR] = 2027
    AND [Month] = 4
    AND [DAY] = 16;

UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'29 Tết Âm lịch 2028'
WHERE [YEAR] = 2028
    AND [Month] = 1
    AND [DAY] = 25;
UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Mùng 1 Tết Âm lịch 2028'
WHERE [YEAR] = 2028
    AND [Month] = 1
    AND [DAY] = 26;

UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Giỗ Tổ Hùng Vương- giải phóng miền Nam 2028'
WHERE [YEAR] = 2028
    AND [Month] = 4
    AND [DAY] = 30;

UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'29 Tết Âm lịch 2029'
WHERE [YEAR] = 2029
    AND [Month] = 2
    AND [DAY] = 12;
UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Mùng 1 Tết Âm lịch 2029'
WHERE [YEAR] = 2029
    AND [Month] = 2
    AND [DAY] = 13;

UPDATE Dim_Date

```

```

SET [IsHoliday] = 1,
    [HolidayName] = N'Giỗ Tổ Hùng Vương 2029'
WHERE [YEAR] = 2029
    AND [Month] = 4
    AND [DAY] = 23;

UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'29 Tết Âm lịch 2030'
WHERE [YEAR] = 2030
    AND [Month] = 1
    AND [DAY] = 31;
UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Mùng 1 Tết Âm lịch 2030'
WHERE [YEAR] = 2030
    AND [Month] = 2
    AND [DAY] = 1;

UPDATE Dim_Date
SET [IsHoliday] = 1,
    [HolidayName] = N'Giỗ Tổ Hùng Vương 2030'
WHERE [YEAR] = 2030
    AND [Month] = 4
    AND [DAY] = 12;

```

Dữ liệu sau khi insert:

	Date_id	Date	Day	DaySuffix	Weekday	WeekDayName	WeekDayName_Short	WeekDayName_FirstLetter	DOWInMonth	DayOfYear	WeekOfMonth	WeekOfYear	Mo
1	20180101	2018-01-01	1	st	2	Monday	MON	M	1	1	1	1	1
2	20180102	2018-01-02	2	nd	3	Tuesday	TUE	T	2	2	1	1	1
3	20180103	2018-01-03	3	rd	4	Wednesday	WED	W	3	3	1	1	1
4	20180104	2018-01-04	4	th	5	Thursday	THU	T	4	4	1	1	1
5	20180105	2018-01-05	5	th	6	Friday	FRI	F	5	5	1	1	1
6	20180106	2018-01-06	6	th	7	Saturday	SAT	S	6	6	1	1	1
7	20180107	2018-01-07	7	th	1	Sunday	SUN	S	7	7	2	2	1
8	20180108	2018-01-08	8	th	2	Monday	MON	M	8	8	2	2	1
9	20180109	2018-01-09	9	th	3	Tuesday	TUE	T	9	9	2	2	1
10	20180110	2018-01-10	10	th	4	Wednesday	WED	W	10	10	2	2	1
11	20180111	2018-01-11	11	th	5	Thursday	THU	T	11	11	2	2	1

Hình 4.2 Kết quả sau khi thực thi câu lệnh Insert bảng Dim_Date

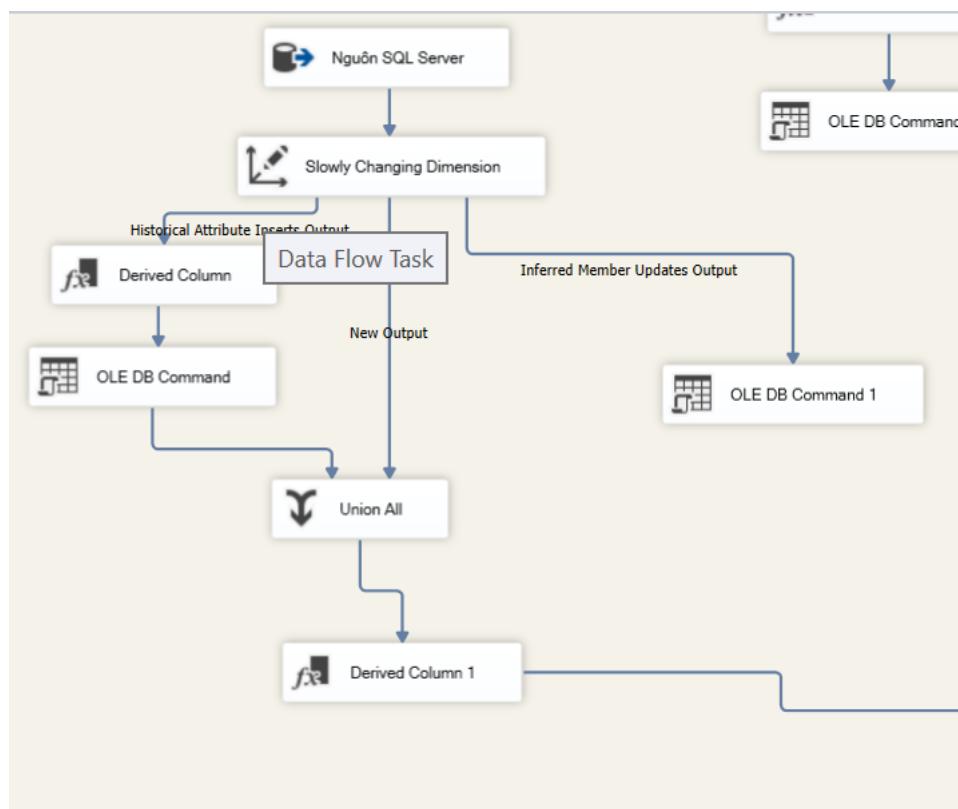
4.3.2. Xây dựng bảng cắt lớp Học viên (Dim_HocVien)

Thông tin học viên là dữ liệu quan trọng tại trung tâm, là master data ảnh hưởng tới tất cả các quy trình nghiệp vụ. Các phòng ban đều có các nghiệp vụ quan trọng với dữ liệu về học viên. Tuy nhiên ở cả 3 phòng dữ liệu học viên chưa đồng nhất.

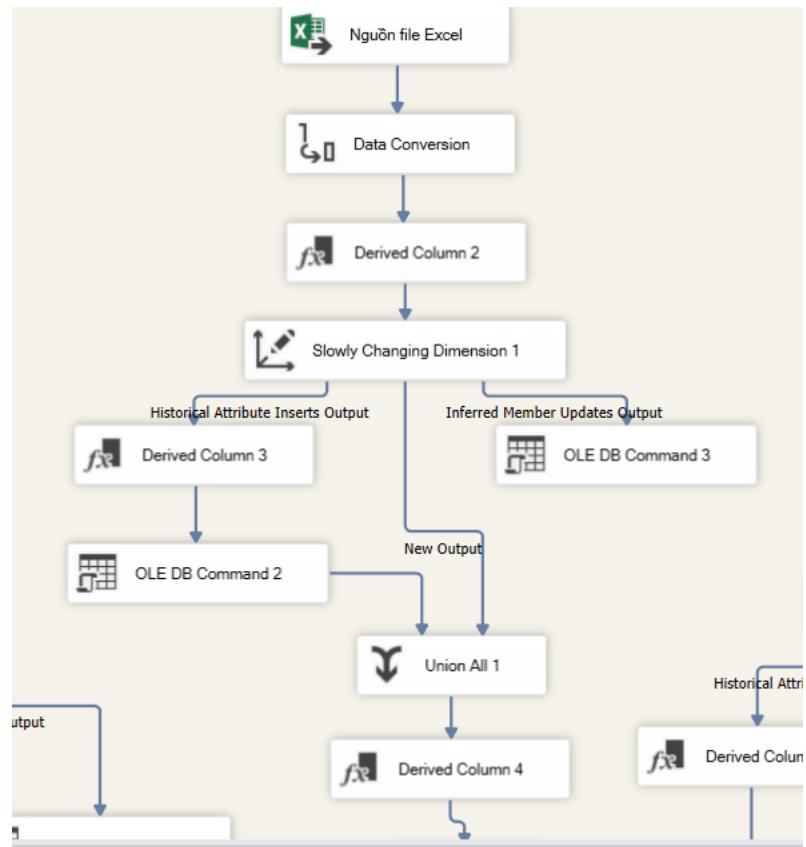
Dim_HocVien *	
ID_HV	
Hoten	
Gioitinh	
Sodt	
Email	
TrinhDo	
Updatedate	
Endate	
CurrentFlag	

Hình 4.3 Bảng Dim_HocVien

- Tại Nguồn 1 (SQL source) dữ liệu là Họ tên, giới tính: Nam/Nữ.
- Tại nguồn 2 (File Excel) dữ liệu là Họ đệm và Tên, giới tính: Male/Female.
- Tại nguồn 3 (MySQL source) dữ liệu là Họ tên và giới tính: 0/1.

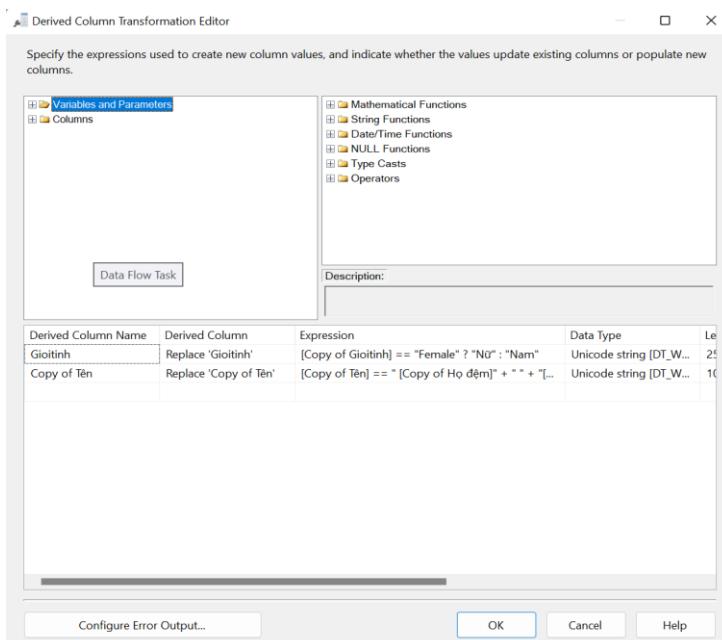


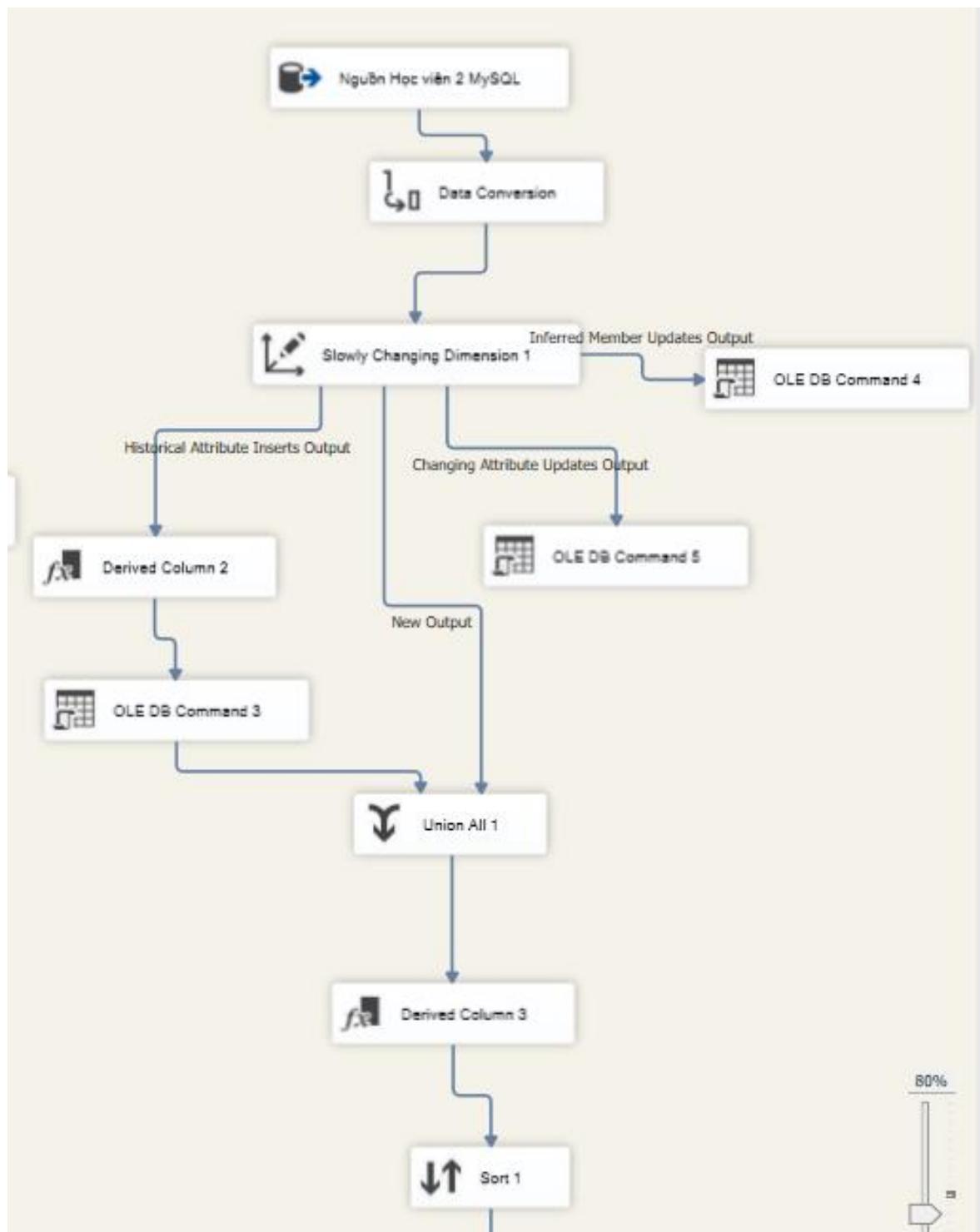
Hình 4.4 Nguồn 1 (SQL Server source)



Hình 4.5 Nguồn 2 (File Excel)

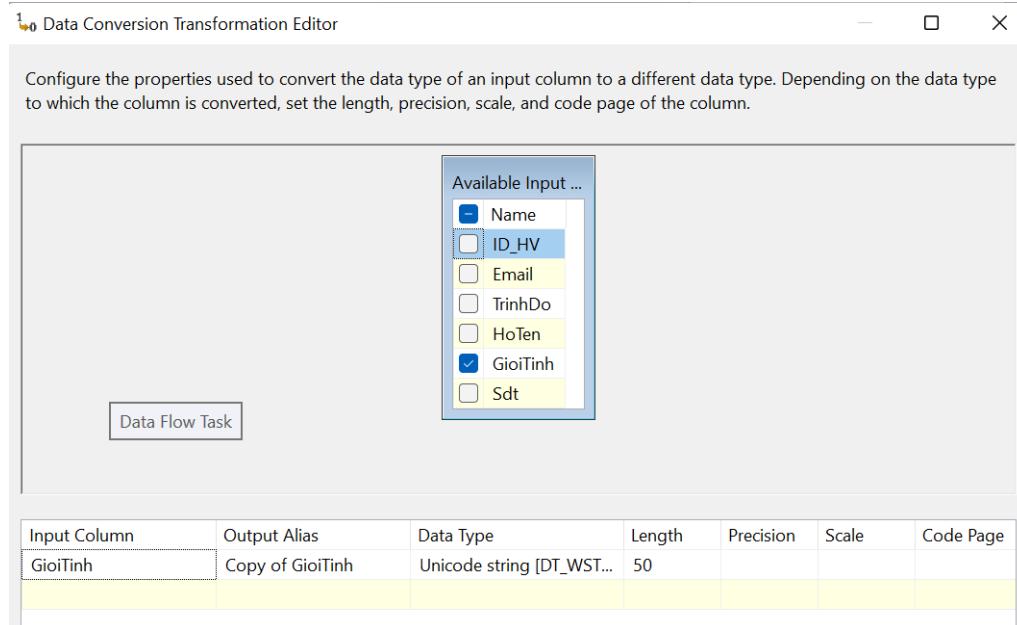
Hình trên mô tả quá trình gộp họ đệm và tên thành họ tên đồng thời đổi giá trị trường “Gioitinh” từ Female/Male thành Nam/Nữ.



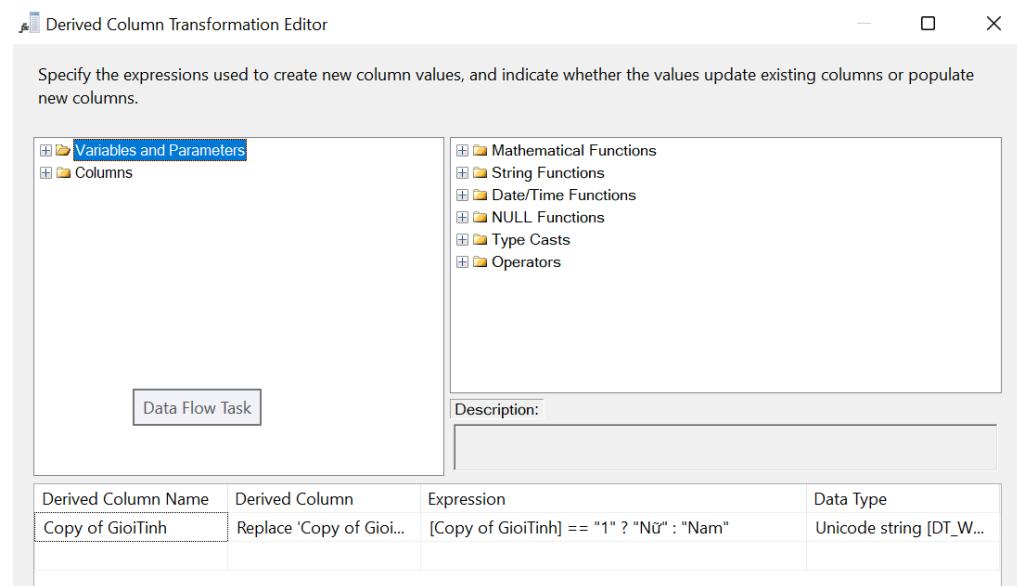


Hình 4.6 Nguồn 3 (MySQL source)

Tại nguồn 3 tiến hành task Data conversion 1: tạo trường copy of GioiTinh và ép sang kiểu Unicode string.

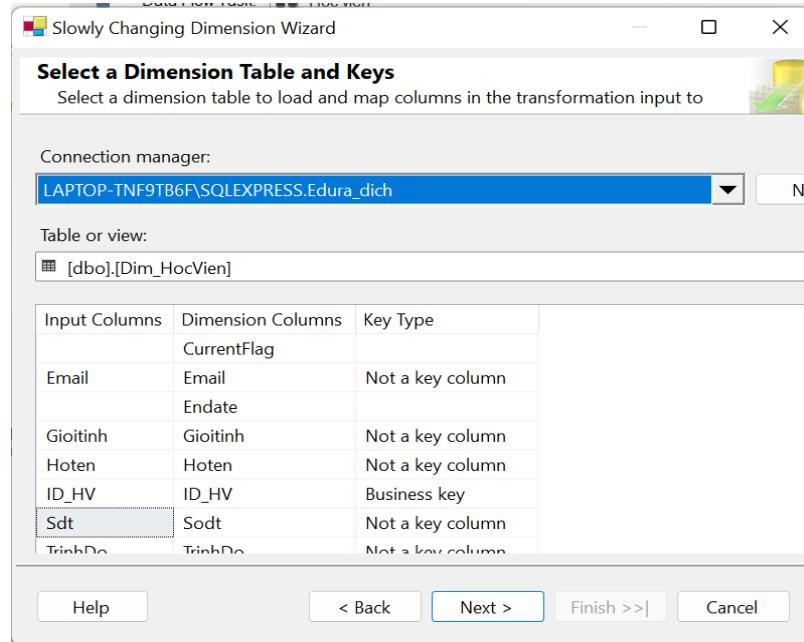


Sau đó thực hiện đổi giới tính từ 0/1 sang Nam/Nữ như hình bên dưới.

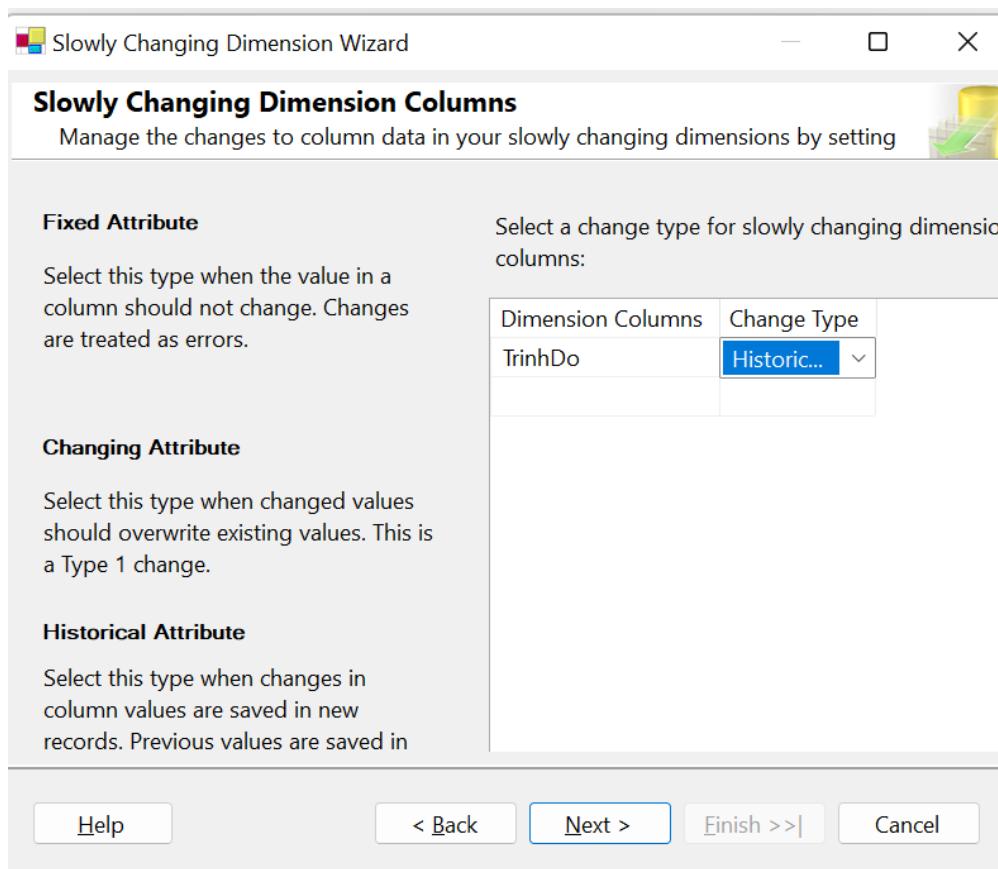


Giá trị cột “TrinhDo” giữa các phòng ban có giá trị là khác nhau. Điều này là do học viên học tại trung tâm giúp cải thiện và tiến bộ, trình độ được nâng cao. Như vậy trình độ của học viên sẽ thay đổi theo thời gian. Để giải quyết điều này, nhóm Cài đặt Slowly Changing Dimension kiểu 03 ở cả 3 nguồn dữ liệu.

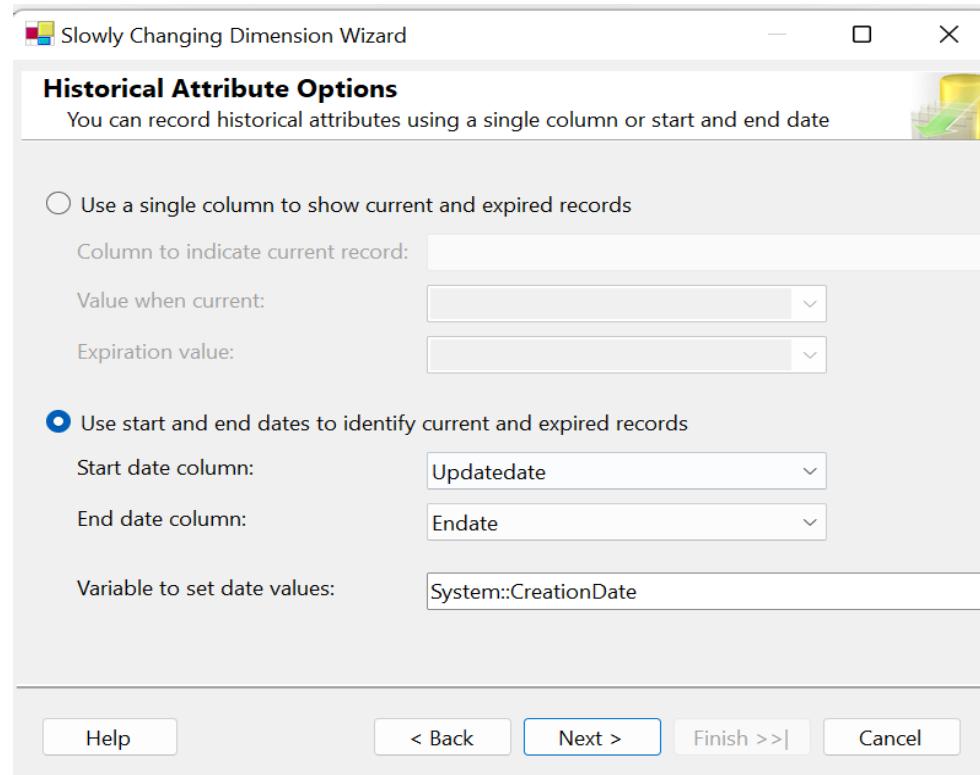
- Chọn bảng Dim_HocVien, chọn Key Type của ID_HV là Business key.



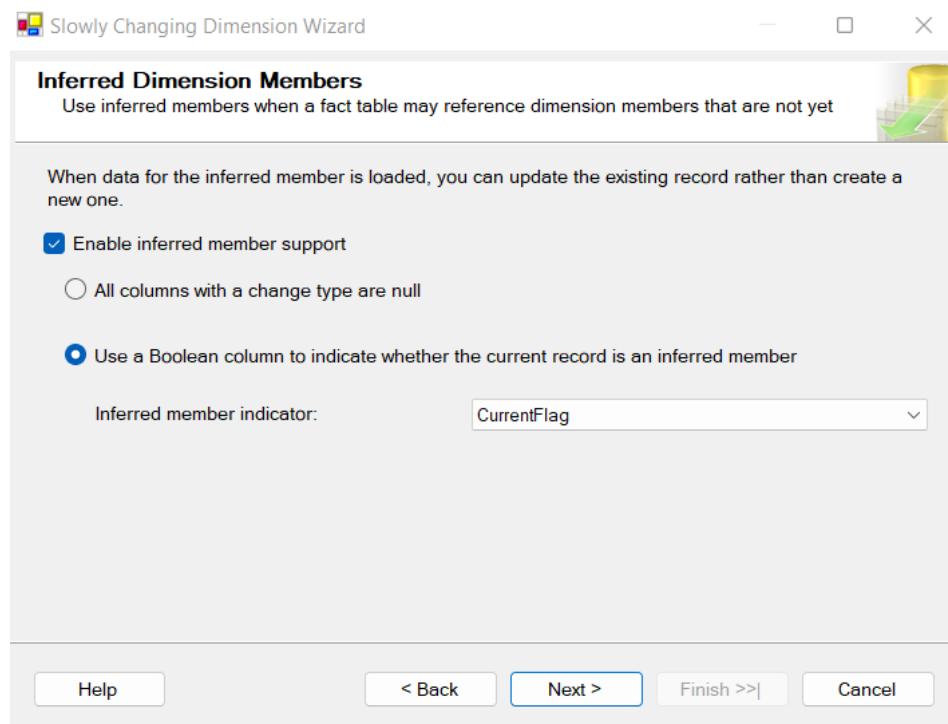
- Trường “Trinhdo” là trường thuộc tính thay đổi có lịch sử.



- Trường *Updatedate* và *Endate* phục vụ việc lưu lại lịch sử thay đổi trường “TrinhDo”.



- Trường *CurrentFlag* được sử dụng để đánh dấu dữ liệu mới hay cũ:



- Sau đó thực hiện merge join 3 nguồn.



Hình 4.7 Thực hiện đỗ dữ liệu từ các nguồn tới bảng Dim_NVSale

Kết quả thu được tại bảng Dim_HocVien:

ID_HV	Hoten	Gioitinh	Sodt	Email	TrinhDo	Updatedate	Endate	CurrentFlag
1	20200001	False	Male	NULL	phantu93@gmail.com	Kaiwa 4	2022-09-18	NULL
2	20200002	False	Male	973523920	huycuongbk0490@gmail.com	Kaiwa 3	2022-09-18	NULL
3	20200003	False	Female	332534001	nguyenthilinh040690@gmail.com	Kaiwa 1	2022-09-18	NULL
4	20200004	False	Female	818066368686	trulythaitran1992@gmail.com	Kaiwa 3	2022-09-18	NULL
5	20200005	False	Female	7022604664	mydung060400@gmail.com	Kaiwa 4	2022-09-18	NULL
6	20200006	False	Female	378378198	duchuot08@gmail.com	Kaiwa 3	2022-09-18	NULL
7	20200007	False	Male	902380306	thienvumai@gmail.com	Kaiwa 3	2022-09-18	NULL
8	20200008	False	Male	972561243	phanvantrixd4a2@gmail.com	Kaiwa 1	2022-09-18	NULL
9	20200009	False	Female	355990660	thuhao160889@gmail.com	Kaiwa 1	2022-09-18	NULL
10	20200010	False	Female	819857793	haiyenkute241193@gmail.com	Kaiwa 1	2022-09-18	NULL
11	20200011	False	Male	968307858	nguyenviettho0211969@gmail.com	Kaiwa 2	2022-09-18	NULL
12	20200012	False	Male	387209577	ducanh1421811@gmail.com	Kaiwa 2	2022-09-18	NULL
13	32020100	False	Male	974534301	mrlamchuc@gmail.com	Kaiwa 3	2022-09-18	NULL
14	42020001	False	Male	NULL	ntanthanh222@gmail.com	Kaiwa 3	2022-09-18	NULL
15	42020002	False	Female	1263296527	buibuidenchap@gmail.com	Kaiwa 2	2022-09-18	NULL
16	42020003	False	Male	979416304	Oovanvuong@gmail.com	Kaiwa 4	2022-09-18	NULL
17	42020004	False	Female	328175791	xanhnt2002@gmail.com	Kaiwa 3	2022-09-18	NULL
18	42020005	False	Female	394439212	yukapham143@gmail.com	Kaiwa 3	2022-09-18	NULL
19	42020006	False	Male	939825668	phuocnguyen20062001@gmail.com	Kaiwa 1	2022-09-18	NULL
20	42020007	False	Female	388619520	ntb2890@gmail.com	Kaiwa 4	2022-09-18	NULL
21	42020008	False	Male	1688288176	Nguyenthanhduy98765@gmail.com	Kaiwa 1	2022-09-18	NULL
22	42020013	False	Male	986764383	tuanduyk100401@gmail.com	Kaiwa 2	2022-09-18	NULL
23	42020014	False	Female	902212869	thanhthiem@hlu.edu.vn	Kaiwa 1	2022-09-18	NULL

Hình 4.8 Kết quả thu được tại bảng Dim_HocVien

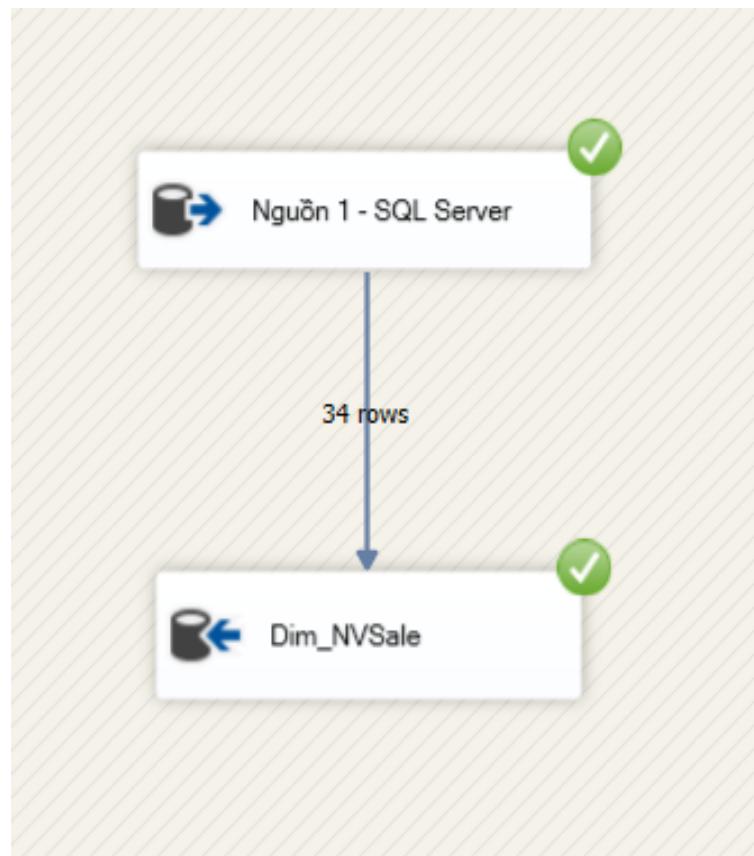
4.3.3. Xây dựng bảng cắt lớp nhân viên Sale (Dim_NVSale)

Bảng cắt lớp nhân viên sale mô các thông tin của nhân viên sale cần lưu trữ như: Mã nhân viên sale, tên nhân viên, giới tính. Điều đó giúp các cấp quản lý (Phòng hành chính nhân sự, ban giám đốc...) dễ dàng kiểm soát KPIs, quản lý lịch trình công việc của nhân sự.

Dim NVSale	
MaNVSale	
Hoten	
Gioitinh	

Hình 4.9 Bảng Dim_NVSale

Thực hiện đổ dữ liệu từ nguồn 1 – Server SQL tới bảng Dim_NVSale.



Hình 4.10 Thực hiện đổ dữ liệu tới bảng Dim_NVSale

	MaNVSale	Hoten	Gioitinh
1	Anhvtl	Vũ Thị Lan Anh	Nữ
2	Daonta	Nguyễn Thị Ánh Đào	Nữ
3	Dinhtc	Trần Công Định	Nam
4	Dungtt1	Trần Thị Dung	Nữ
5	Duyenlt	Lê Thu Duyên	Nữ
6	Hoaidtt	Đặng Thị Thanh Hoài	Nữ
7	Huongnt	Nguyễn Thị Hướng	Nữ
8	Huongnt1	Nguyễn Thu Hướng	Nữ
9	Huongntt	Nguyễn Thị Thu Hướng	Nữ
10	Huongvtl	Vũ Thị Lan Hướng	Nữ
11	huyenlm	Lê Minh Huyền	Nữ
12	Huyenptk	Phan Thị Khánh Huyền	Nữ
13	Khuent	Nguyễn Thanh Khuê	Nữ
14	Kimdtt	Đặng Thị Thanh Kim	Nữ
15	Lanttt	Trần Thị Thu Lan	Nữ
16	Linhntm	Nguyễn Thị Mỹ Linh	Nữ
17	Linhtk	Trần Khánh Linh	Nữ
18	Linhtna	Trần Nguyễn Ánh Linh	Nữ
19	Longph	Phan Hoàng Long	Nam
20	Minhha	Hà Ánh Minh	Nữ

Hình 4.11 Kết quả thu được tại bảng Dim_NVSale

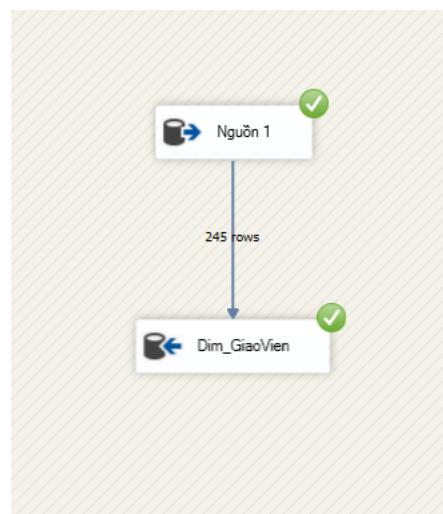
4.3.4. Xây dựng bảng cắt lớp giáo viên (Dim_GiaoVien)

Bảng Dim_GiaoVien mô tả các thông tin như: mã giáo viên, tên giáo viên, quốc tịch, email, giới tính, mức lương để các phòng ban có thể dễ dàng quản lý.

Dim GiaoVien	
▼	ID_GV
	Hoten
	Quoctich
	Email
	Gioitinh
	Mucluong

Hình 4.12 Bảng Dim_GiaoVien

Dữ liệu bảng Dim_GiaoVien được lấy từ nguồn 1 (SQL source).



Hình 4.13 Thực hiện đổ dữ liệu từ nguồn 1 tới bảng Dim_GiaoVien

	ID_GV	Hoten	Quoclich	Email	Gioitinh	Mucluong
1	GV01	Ichinose Miho	Nhật Bản	iamicns01@gmail.com	Nữ	950
2	GV02	Nakao Yuta	Nhật Bản	sakuratoso.nakao@gmail.com	Nữ	950
3	GV03	Chiharu Yabuki	Nhật Bản	chiharu458@gmail.com	Nữ	950
4	GV04	Phan Nhung	Việt Nam	Phannhungv97@gmail.com	Nữ	120000
5	GV05	Kamiunten Hiroko	Nhật Bản	heroxxkamiunten@gmail.com	Nữ	950
6	GV06	Miyata Mariko	Nhật Bản	t.mariko.shun@gmail.com	Nữ	950
7	GV07	Đinh Văn Hoàng	Việt Nam	hoangdinh.tohoku@gmail.com	Nam	150000
8	GV08	Đặng Thị Thu Hà	Việt Nam	dangha94bg@gmail.com	Nữ	120000
9	GV09	Võ Chí Thiên	Việt Nam	chithien1993spk@gmail.com	Nam	120000
10	GV10	Kinugasa Moe	Nhật Bản	pro.kinugasa@gmail.com	Nam	950
11	GV100	Homma YUI	Nhật Bản	yui.2hongo@gmail.com	Nữ	950
12	GV101	Quang Thị Phương	Việt Nam	quangphuong1410@gmail.com	Nữ	105000
13	GV102	Miyako Ayako	Nhật Bản	ayaco_1224@yahoo.co.jp	Nữ	950
14	GV103	Đặng Thị Dương	Việt Nam	duong76.hanu.jp@gmail.com	Nữ	105000
15	GV104	Nguyễn Ngọc Sá...	Việt Nam	ngocsangwind@gmail.com	Nữ	110000
16	GV105	Tran Thi My Linh	Việt Nam	mylinh291995@gmail.com	Nữ	100000
17	GV106	Trần Hoàng Hiệp	Việt Nam	hieplhp99@gmail.com	Nam	100000
18	GV107	Nguyễn Trịnh Tú ...	Việt Nam	tulinhvnpj@gmail.com	Nữ	110000
19	GV108	Shimoda Keito	Nhật Bản	pop.keito@icloud.com	Nam	950
20	GV109	Moriyama Saho	Nhật Bản	junjun.smyluv4444@gmail.com	Nữ	1000
21	GV11	Trần Linh Chi	Việt Nam	linhchitran18@gmail.com	Nữ	130000
22	GV110	Nguyễn Tiến Mạnh	Việt Nam	nguyentienmanh26192@gma...	Nam	105000

Query executed successfully. | LAPTOP-TH

Hình 4.14 Kết quả thu được tại bảng Dim_GiaoVien

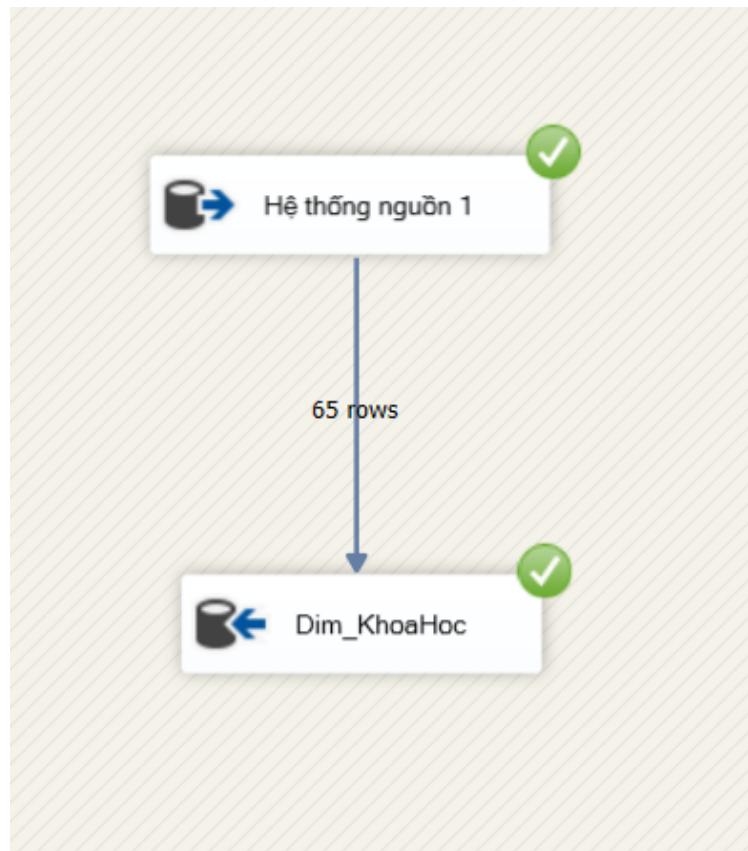
4.3.5. Xây dựng bảng cắt lớp khóa học (Dim_KhoaHoc)

Bảng Dim_KhoaHoc cung cấp thông tin về mã khóa học, tên khóa học, số buổi của khóa học, loại khóa học (1 giáo viên – 1 học viên hoặc 1 giáo viên nhiều học viên), học phí và giá khuyến mại tùy vào chương trình khuyến mãi. Từ đó giúp các xây dựng các báo cáo doanh thu theo từng tháng năm quý....

Dim KhoaHoc	
🔑	ID_KH
Tên	Sobuoi
Loaikhoa	Loaikhoa
Hocphi	Hocphi
GiaKM	GiaKM

Hình 4.15 Bảng Dim_KhoaHoc

Dữ liệu bảng Dim_KhoaHoc được lấy từ nguồn 1 (SQL source).



Hình 4.16 Thực hiện đổ dữ liệu cho bảng Dim_KhoaHoc

	ID_KH	Tên	Sobuoi	Loaikhoa	Hocphi	GiaKM
1	CKAI11221030	Combo_kaiwa12VN	60	1-1	20875000	14500000
2	CKAI121030JP	combo_kaiwa12	60	1-1	22625000	15500000
3	CKAI121030VN	combo_kaiwa12	60	1-1	19000000	11500000
4	CKAI12331030	Combo_kaiwa12V3N	90	1-1	31500000	21500000
5	CKAI12341030	Combo_kaiwa12V34N	120	1-1	44625000	34000000
6	CKAI12351030	Combo_kaiwa12V345N	150	1-1	57750000	36000000
7	CKAI12N060VN	Combo_BS12V	60	1-4	16000000	6500000
8	CKAI131030JP	combo_kaiwa123	90	1-1	35125000	23500000
9	CKAI131030VN	combo_kaiwa123	90	1-1	30000000	17000000
10	CKAI13N060VN	Combo_BS123N	90	1-4	25500000	9500000
11	CKAI141030JP	combo_kaiwa1234	120	1-1	48250000	32000000
12	CKAI141030VN	combo_kaiwa1234	120	1-1	43125000	23500000
13	CKAI14N060VN	Combo_BS1234N	120	1-4	35500000	12500000
14	CKAI151030JP	combo_kaiwa12345	150	1-1	61375000	40000000
15	CKAI151030VN	combo_kaiwa12345	150	1-1	56250000	30500000
16	CKAI15N060VN	Combo_BS12345N	150	1-4	46500000	15500000
17	CKAI22331030	Combo_kaiwa23VN	60	1-1	21500000	15500000
18	CKAI22341030	Combo_kaiwa2V34N	90	1-1	35625000	24000000
19	CKAI22351030	Combo_kaiwa2V345N	120	1-1	48750000	32500000

Query executed successfully.

Hình 4.17 Kết quả thu được tại bảng Dim_KhoaHoc

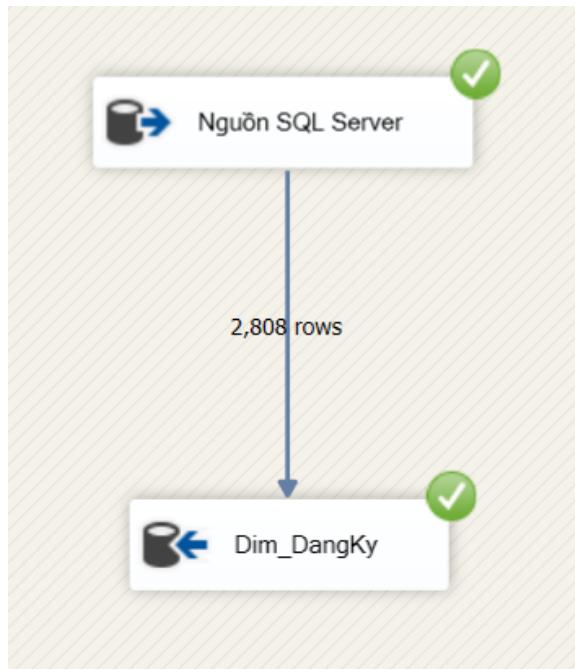
4.3.6. Xây dựng bảng cốt lõi Đăng Ký (Dim_DangKy)

Bảng Dim_Dangky cung cấp thông tin về mã học viên, mã nhân viên sale, ngày đăng ký khóa học. Sử dụng những nguồn thông tin này có thể để thống kê, đánh giá số lượng học viên đăng ký qua mỗi ngày, mỗi đợt, qua đó có những chính sách đặc biệt vào những ngày như vậy để áp dụng những chính sách đặc biệt vào những tháng, năm kế tiếp qua đó tăng doanh số và đem về nhiều hơn lợi nhuận cho doanh nghiệp.

Dim DangKy	
ID_HV	
MaNVSale	
ID_KH	
Ngaydk	

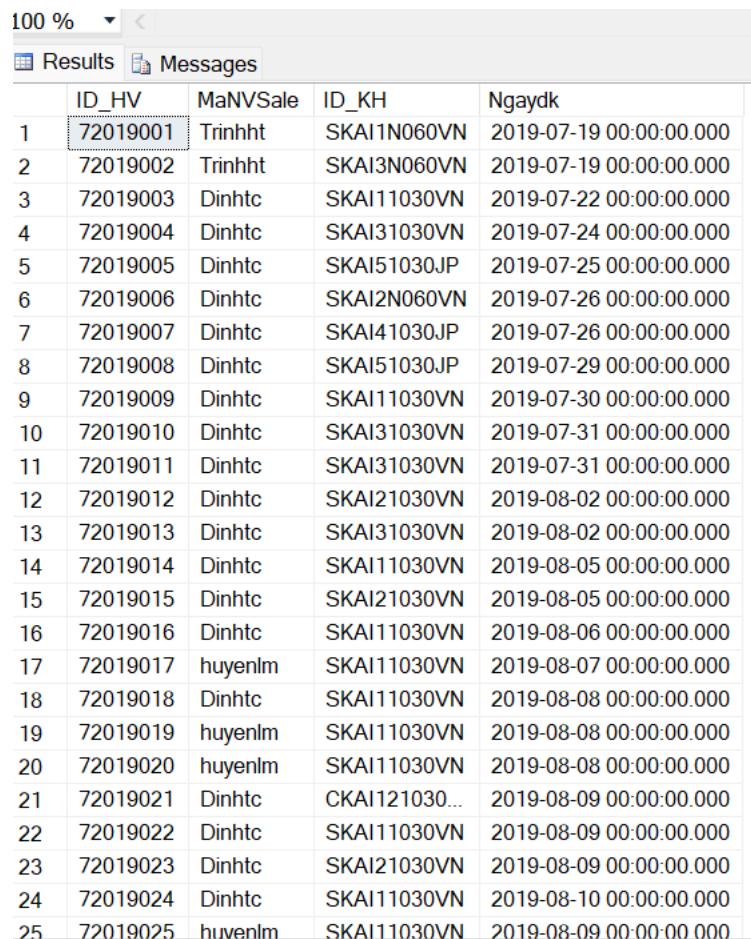
Hình 4.18 Bảng Dim_DangKy

Dữ liệu bảng Dim_Dangky được lấy từ nguồn 1 (SQL source).



Hình 4.19 Thực hiện đổ dữ liệu tới bảng Dim_DangKy

Kết quả thu được tại bảng Dim_DangKy:



	ID_HV	MaNVSale	ID_KH	Ngaydk
1	72019001	Trinhht	SKAI1N060VN	2019-07-19 00:00:00.000
2	72019002	Trinhht	SKAI3N060VN	2019-07-19 00:00:00.000
3	72019003	Dinhtc	SKAI11030VN	2019-07-22 00:00:00.000
4	72019004	Dinhtc	SKAI31030VN	2019-07-24 00:00:00.000
5	72019005	Dinhtc	SKAI51030JP	2019-07-25 00:00:00.000
6	72019006	Dinhtc	SKAI2N060VN	2019-07-26 00:00:00.000
7	72019007	Dinhtc	SKAI41030JP	2019-07-26 00:00:00.000
8	72019008	Dinhtc	SKAI51030JP	2019-07-29 00:00:00.000
9	72019009	Dinhtc	SKAI11030VN	2019-07-30 00:00:00.000
10	72019010	Dinhtc	SKAI31030VN	2019-07-31 00:00:00.000
11	72019011	Dinhtc	SKAI31030VN	2019-07-31 00:00:00.000
12	72019012	Dinhtc	SKAI21030VN	2019-08-02 00:00:00.000
13	72019013	Dinhtc	SKAI31030VN	2019-08-02 00:00:00.000
14	72019014	Dinhtc	SKAI11030VN	2019-08-05 00:00:00.000
15	72019015	Dinhtc	SKAI21030VN	2019-08-05 00:00:00.000
16	72019016	Dinhtc	SKAI11030VN	2019-08-06 00:00:00.000
17	72019017	huyenlm	SKAI11030VN	2019-08-07 00:00:00.000
18	72019018	Dinhtc	SKAI11030VN	2019-08-08 00:00:00.000
19	72019019	huyenlm	SKAI11030VN	2019-08-08 00:00:00.000
20	72019020	huyenlm	SKAI11030VN	2019-08-08 00:00:00.000
21	72019021	Dinhtc	CKAI121030...	2019-08-09 00:00:00.000
22	72019022	Dinhtc	SKAI11030VN	2019-08-09 00:00:00.000
23	72019023	Dinhtc	SKAI21030VN	2019-08-09 00:00:00.000
24	72019024	Dinhtc	SKAI11030VN	2019-08-10 00:00:00.000
25	72019025	huyenlm	SKAI11030VN	2019-08-09 00:00:00.000

Hình 4.20 Kết quả thu được tại bảng Dim_DangKy

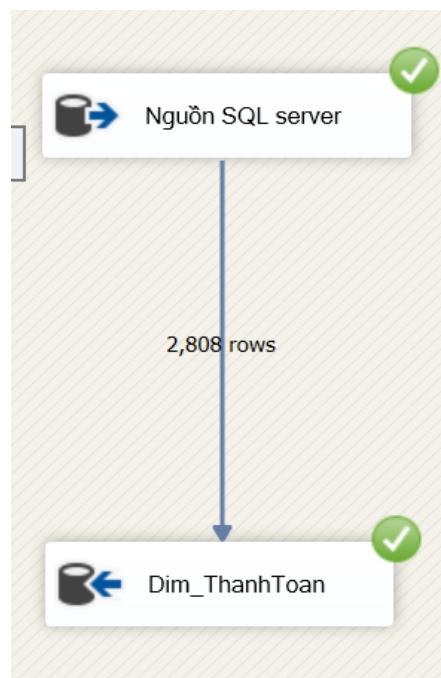
4.3.7. Xây dựng bảng cắt lớp Thanh Toán (Dim_ThanhToan)

Bảng Dim_ThanhToan mô tả thông tin về thống kê các vấn đề liên quan đến tình hình tài chính của học viên với tổ chức Edura. Bảng Dim_ThanhToan bao gồm những dữ liệu mã học viên, mã khóa học, trạng thái thanh toán, ngân hàng nhận, số tiền Nhật phải nộp, số tiền Việt phải nộp. Từ những thông tin đó làm căn cứ để công ty xác định được những học viên tiềm năng hay đăng ký khóa học, những học viên còn chậm hay chưa thanh toán học phí qua đó giúp tổ chức/ doanh nghiệp có những chiến lược, chính sách, nhắc nhở cụ thể với từng học viên học tập ở công ty Edura.

Dim_ThanhToan	
ID_HV	
ID_KH	
TrangthaiTT	
NgayTT	
[Nganhangnhan (KT)]	
Sotiennop_Nhat	
Sotiennop_Viet	

Hình 4.21 Bảng Dim_ThanhToan

Dữ liệu bảng Dim_ThanhToan được lấy từ nguồn 1 (SQL source).



Hình 4.22 Thực hiện đổ dữ liệu tới bảng Dim_ThanhToan

	ID_HV	ID_KH	TrangthaiTT	NgayTT	Nganhangnhan (KT)	Sotiennop_Nhat	Sotiennop_Viet
1	420234761	CKAI121030VN	NULL	2021-11-02 00:00:00.000	Vietcombank (0041000142825)	NULL	12500000
2	420234547	CKAI231030VN	Full phí	2021-11-02 00:00:00.000	Vietcombank (0041000142825)	NULL	13500000
3	420235005	CKAI231030VN	Full phí	2021-11-04 00:00:00.000	MB (6660914491211)	64676	13030000
4	420234256	SKAI21030VN	Đặt cọc	2021-10-27 00:00:00.000	MB (6660914491211)	10000	2000000
5	420234769	CKAIP211030VN	Full phí	2021-11-05 00:00:00.000	Vietcombank (0041000142825)	NULL	16000000
6	420235050	SKAI21030VN	Full phí	2021-11-06 00:00:00.000	MB (6660914491211)	35000	7034000
7	420234896	CKAI231030VN	Full phí	2021-11-05 00:00:00.000	Vietcombank (0041000142825)	NULL	12500000
8	420235196	SKAI21030VN	Full phí	2021-11-05 00:00:00.000	Vietcombank (0041000142825)	NULL	7000000
9	420235232	SKAI31030JP	Thanh toán lần 1	2021-11-05 00:00:00.000	MB (6660914491211)	22500	4500000
10	420234631	SKAI21030VN	Full phí	2021-11-05 00:00:00.000	MB (6660914491211)	35000	7035000
11	420235111	SKAI21030VN	Full phí	2021-11-05 00:00:00.000	MB (6660914491211)	35000	7035000
12	420234916	SKAI21030VN	Thanh toán lần 1	2021-11-07 00:00:00.000	Vietcombank (0041000142825)	NULL	5000000
13	420235146	SKAI21030VN	Thanh toán lần 1	2021-11-07 00:00:00.000	MB (6660914491211)	21000	4200000
14	420235193	SKAI31030VN	Full phí	2021-11-08 00:00:00.000	Vietcombank (0041000142825)	NULL	8000000
15	420235147	SKAI31030JP	Full phí	2021-11-07 00:00:00.000	Vietinbank (103870317215)	NULL	10500000
16	420234963	CKAI231030VN	NULL	2021-11-07 00:00:00.000	Vietcombank (0041000142825)	NULL	8000000
17	420234370	CKAI33441030	Thanh toán lần 1	2021-11-08 00:00:00.000	MB (6660914491211)	50000	10050000
18	420235048	CKAI341030JP	Thanh toán lần 1	2021-11-08 00:00:00.000	Vietcombank (0041000142825)	NULL	10000000
19	420235242	CKAIP211030VN	Full phí	2021-11-08 00:00:00.000	Vietcombank (0041000142825)	NULL	16000000

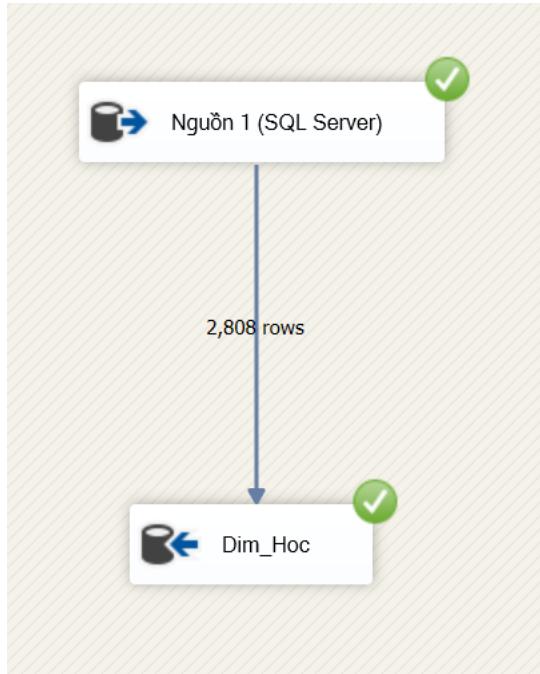
Query executed successfully.

Hình 4.23 Kết quả thu được tại bảng Dim_ThanhToan

4.3.8. Xây dựng bảng cắt lớp Học (Dim_Hoc)

Bảng Dim_Hoc với đầy đủ các dữ liệu liên quan đến việc học của học viên mà công ty đang kinh doanh với những trường dữ liệu mã học viên, mã khóa học, mã giáo viên, trạng thái lớp học, ngày bắt đầu khóa học, ngày kết thúc khóa học, ngày học, tổng số buổi học, số buổi đã học, số buổi còn lại. Bảng Dim_Hoc sẽ giúp bạn quản lý, phụ trách các học và các giáo viên của công ty Edura dễ dàng nắm bắt được tình trạng của các học viên và qua đó kịp thời có những đánh giá, nhắc nhở học viên về tiến trình, kế hoạch của khóa học mình tham gia.

Dữ liệu bảng Dim_Hoc được lấy từ nguồn 1 (SQL source).



Hình 4.24 Thực hiện đỗ dữ liệu tới bảng Dim_Hoc

	ID_HV	ID_KH	ID_GV	Trangthailop	NgayBD	NgayKT_BL	Ngayhoc	Tongboboui	Sobuoiconlai	Sobuidahoc
4	420227854	CKAI121030VN	GV04	Kết thúc	2021-06-18 00:00:00.000	2022-01-12 00:00:00.000	T2-T4-T6	62	-1	63
5	420228205	CKAIP111030VN	GV07	Kết thúc	2021-06-18 00:00:00.000	2021-09-10 00:00:00.000	23456	52	1	51
6	420227945	CKAI231030VN	GV08	Bảo lưu	2021-06-16 00:00:00.000	2021-09-18 00:00:00.000	T2-T4-T6	62	30	32
7	420228195	CKAI451030JP	GV206	Kết thúc	2021-06-18 00:00:00.000	2022-01-09 00:00:00.000	T2-T4-T6	62	-1	63
8	420227959	CKAIP211030VN	GV09	Kết thúc	2021-06-16 00:00:00.000	2021-12-08 00:00:00.000	T2-T4-T6	47	0	47
9	420228200	CKAI121030VN	GV268	Bảo lưu	2021-06-18 00:00:00.000	2021-10-18 00:00:00.000	T2-T4-T6	62	19	43
10	420228285	CKAIP111030VN	GV269	Kết thúc	2021-06-18 00:00:00.000	2021-10-16 00:00:00.000	T2-T4-T6	50	0	50
11	420226885	SKAI11030VN	GV270	Kết thúc	2021-06-18 00:00:00.000	2021-07-10 00:00:00.000	T2-T4-T6	30	9	21
12	420227868	SKAI31030JP	GV159	Kết thúc	2021-06-19 00:00:00.000	2021-08-27 00:00:00.000	T3-T5	32	0	32
13	420228412	SKAI11030VN	GV263	Kết thúc	2021-06-22 00:00:00.000	2021-09-25 00:00:00.000	T3-T5-T7	30	0	30
14	420228473	SKAI11030VN	GV265	Kết thúc	2021-06-23 00:00:00.000	2021-08-02 00:00:00.000	T2-T4	30	19	11
15	420228346	CKAIP211030VN	GV266	Kết thúc	2021-06-22 00:00:00.000	2021-10-29 00:00:00.000	T3-T5-T7	47	0	47

Hình 4.25 Kết quả thu được tại bảng Dim_Hoc

4.3.9. Xây dựng bảng Fact_QLHoc

Bảng fact sales sẽ bao gồm khóa chính của các bảng Dim và các trường như mong muốn của nhà quản lý doanh nghiệp. Bảng fact sales sẽ bao gồm các trường dữ liệu sau: mã học viên, mã khóa học, mã giáo viên, trạng thái lớp học, ngày bắt đầu, ngày kết thúc, ngày học, tổng số buổi học, số buổi còn lại, trạng thái thanh toán, ngày đăng ký mã thời gian. Bởi vì thông thường nhà quản lý sẽ đưa ra các mong muốn hỏi như thông tin liên quan đến tình trạng của học viên cụ thể như ngày học, thanh toán chưa hay số buổi học còn lại nên bảng Fact sẽ có các trường dữ liệu rõ ràng để có thể xuất hiện ở bảng fact.

Fact QLHoc	
ID_HV	
ID_KH	
ID_GV	
Trangthailop	
NgayBD	
NgayKT_BL	
Ngayhoc	
Tongbobuoi	
Sobuoiconlai	
TrangthaiTT	
Ngaydk	
Date_id	

Hình 4.26 Bảng Fact_QLHoc

Ý nghĩa của các trường dữ liệu:

Cột	Data type	Type	Mô tả
ID_HV	float	Dimension	Khóa ngoại bảng Dim_HocVien
ID_KH	varchar(30)	Dimension	Khóa ngoại bảng Dim_KhoaHoc
ID_GV	varchar(20)	Dimension	Khóa ngoại bảng Dim_GiaoVien
Trangthailop	nvarchar(50)	Additive	Trạng thái lớp học
NgayBD	datetime	Additive	Ngày bắt đầu khóa học
NgayKT_BL	datetime	Additive	Ngày kết thúc khóa học
Ngayhoc	nvarchar(100)	Additive	Ngày phải học

Tongsobuoihoc	float	Additive	Tổng số buổi học
Sobuoiconlai	float	Additive	Số buổi còn lại học
TrangthaiTT	nvarchar(50)	Additive	Trạng thái thanh toán
Ngaydk	datetime	Additive	Ngày đăng ký
Date_id	int	Dimension	Khóa ngoại bảng Dim_Date

Bảng 4.1 Các trường dữ liệu của bảng Fact_QLHoc

4.3.10. Xây dựng bảng Fact_Sales

Trong quá trình hoạt động nhà quản lý sẽ có những yêu cầu, đưa ra những mong muốn khóa học được bán theo tuần, tháng, quý nào cho doanh thu cao nhất, nhân viên sales nào đang làm việc hiệu quả với số lượng học viên thu hút được lượng học viên tham gia đăng ký lớn...khi đó các thông tin cần được trả lời sẽ được biểu diễn, thể hiện trên bảng fact sales. Bảng fact Sales sẽ gồm khóa chính của các bảng dim và các trường thông tin mong muốn của nhà quản lý. Bảng fact Sales gồm những trường dữ liệu sau: mã học viên, mã khóa học, trạng thái thanh toán, ngày thanh toán, ngân hàng nhận, số tiền nộp, ngày đăng ký, mã nhân viên sale, ngày nghỉ và mã thời gian.

Fact_Sales	
ID_HV	
ID_KH	
TrangthaiTT	
NgayTT	
[Nganhanganhan (KT)]	
Sotiennop_Nhat	
Sotiennop_Viet	
Ngaydk	
MaNVSale	
IsHoliday	
Date_id	

Hình 4.27 Bảng Fact_Sales

Ý nghĩa của các trường dữ liệu

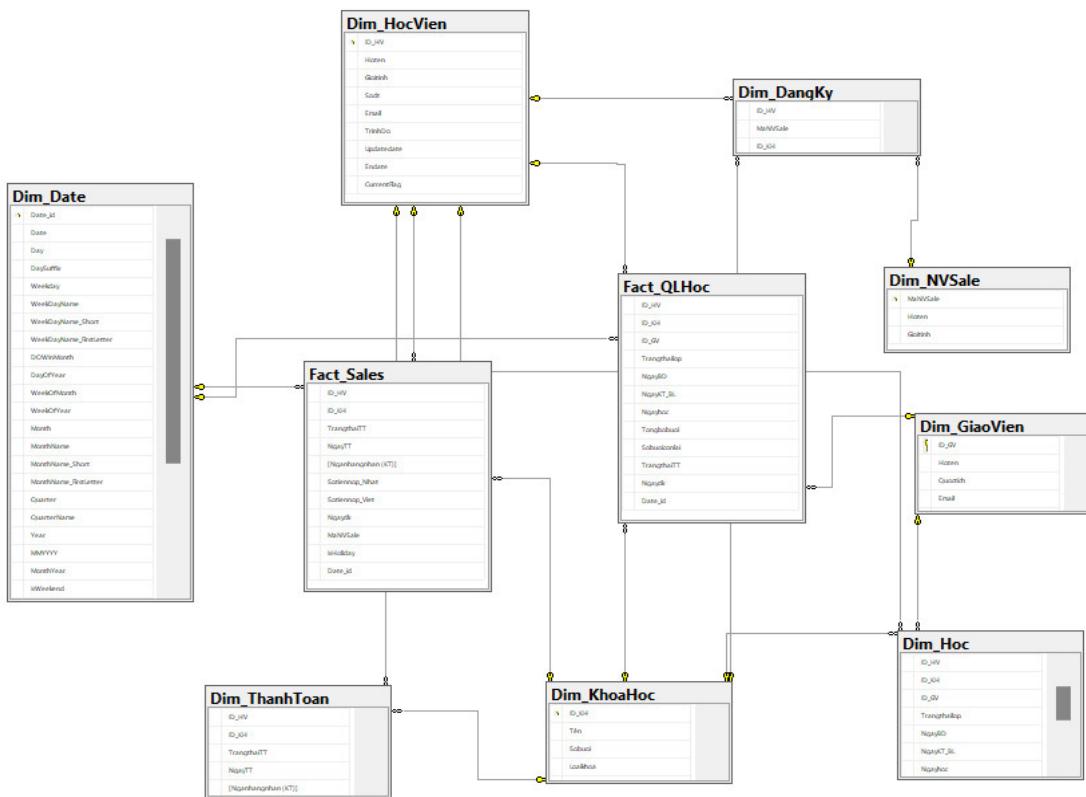
Cột	Data types	Types	Mô tả
ID_HV	float	Dimension	Khóa ngoại bảng Dim_HocVien
ID_KH	varchar(30)	Dimension	Khóa ngoại bảng Dim_KhoaHoc
TrangthaiTT	nvarchar(50)	Additive	Trạng thái thanh toán
NgayTT	datetime	Additive	Ngày thanh toán
Nganhanganhan	nvarchar(255)	Additive	Ngân hàng nhận
Sotiennop_Nhat	float	Additive	Số tiền nộp(Nhật)
Sotiennop_Viet	float	Additive	Số tiền nộp (Việt)
Ngaydk	datetime	Additive	Ngày đăng ký
MaNVsale	varchar(20)	Additive	Mã nhân viên sale

IsHoliday	bit	Additive	Ngày lễ/ thường
Date_id	int	Dimension	Khóa ngoại bảng Dim_Date

Bảng 4.2 Các trường dữ liệu của bảng Fact_Sales

4.3.11. Xây dựng sơ đồ chòm sao

Sơ đồ chòm sao được xây dựng từ bảng Fact và các bảng Dim như sau:

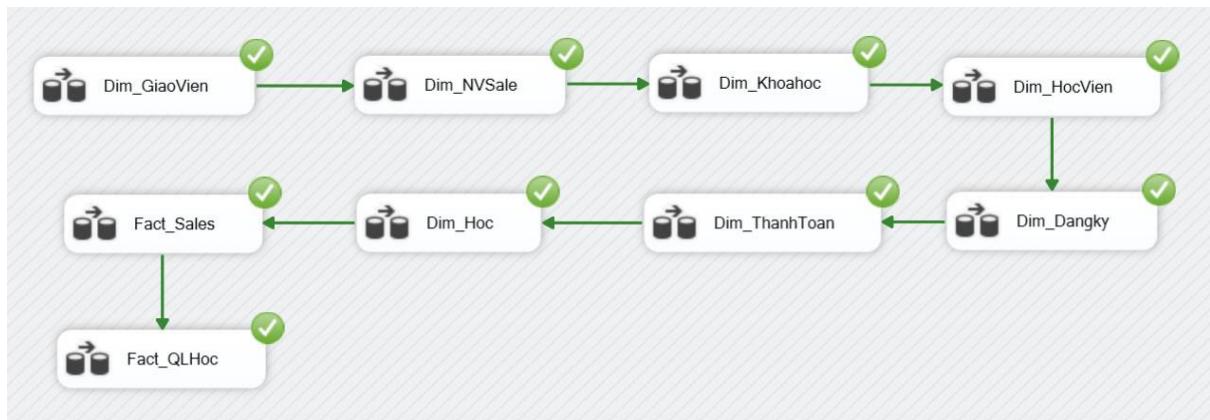


Hình 4.28 Sơ đồ chòm sao

4.4. Cập nhật dữ liệu vào kho dữ liệu

4.4.1. Xây dựng flow

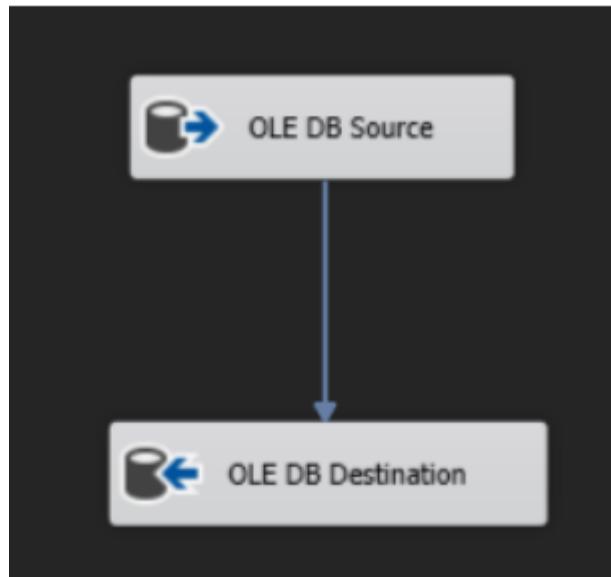
Chương trình cập nhật được xây dựng trên project SSIS, sử dụng trên Visual Studio 2019. Flow chung cập nhật của toàn kho dữ liệu như sau:



Hình 4.29 Data Flow

4.4.2. Data flow bảng Dim_NVsales, Dim_GiaoVien, Dim_KhoaHoc, Dim_DangKy, Dim_ThanhToan, Dim_Hoc

Các bảng này đều được lấy từ nguồn 1 và được thiết kế các cột giống nguồn 1 nên data flow đơn giản, chỉ có 1 nguồn và 1 đích sau đó mappings các cột như sau:

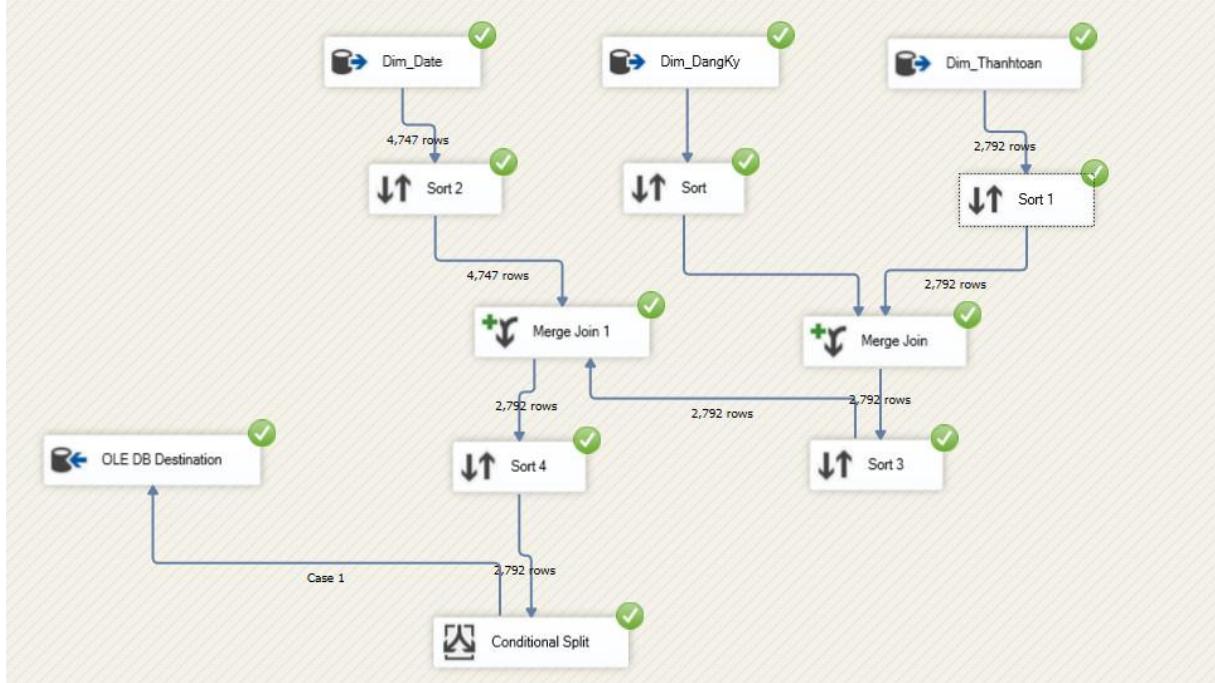


Hình 4.30 Data flow bảng Dim_NVsales, Dim_GiaoVien, Dim_KhoaHoc, Dim_DangKy, Dim_ThanhToan, Dim_Hoc

4.4.3. Data flow bảng Fact_Sales

Bảng Fact_Sales có được dữ liệu bằng cách từ Dim_ThanhToan sau đó Merge Join cùng Dim_DangKy sau đó Merge Join với Dim_Date.

Data flow được xây dựng như sau:



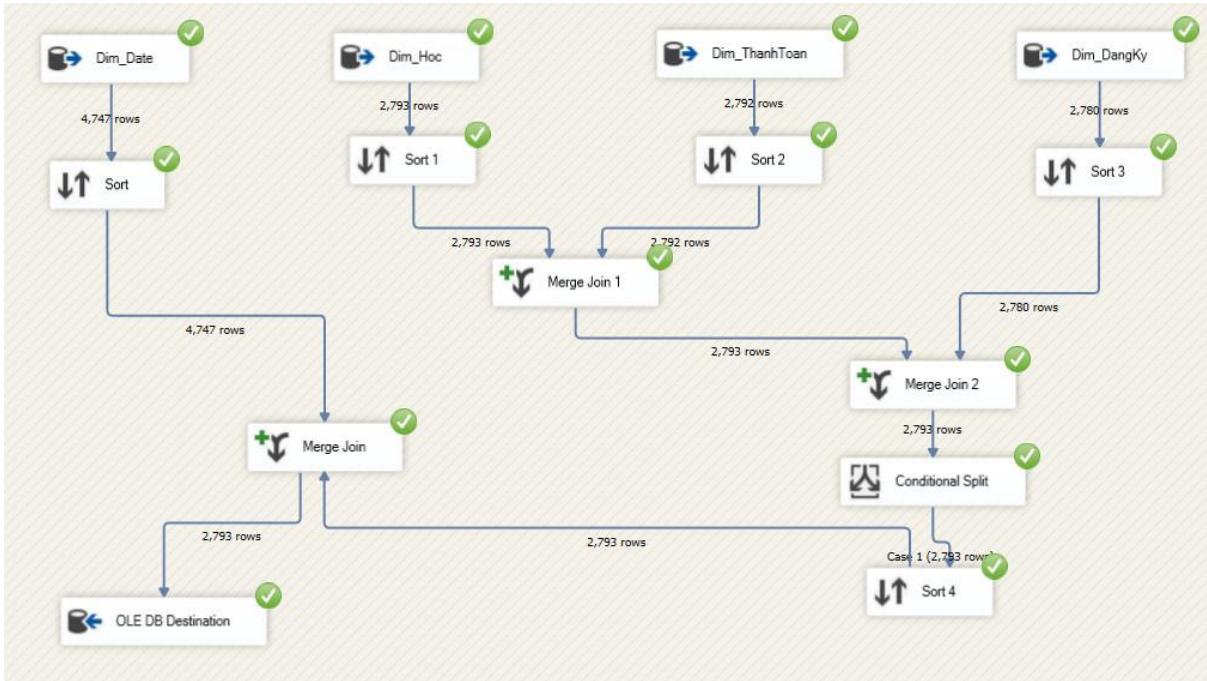
Hình 4.31 Data flow bảng Fact_Sales

ID_HV	ID_KH	TrangthaiTT	NgayTT	Nganhangnhan (KT)	Sotiennop_Nhat	Sotiennop_Viet	Ngaydk	MaNVSale	IsHoliday	Date_id
1	20200001	SKAI41030JP	Thanh toán lần 1	2020-04-04 00:00:00.000	Nhật 1 (Mai - 10160-68856611)	30000	6450000	2020-04-02 00:00:00.000	Anhvtl	0
2	20200002	SKAI31030JP	Full phí	2020-04-04 00:00:00.000	Vietcombank (0041000142825)	NULL	7500000	2020-04-02 00:00:00.000	Trangdtl	0
3	20200003	SKAI11030VN	Full phí	2020-04-06 00:00:00.000	Vietcombank (0041000142825)	NULL	5400000	2020-04-03 00:00:00.000	Trangdtl	0
4	20200004	SKAI31030JP	Full phí	2020-04-06 00:00:00.000	Vietcombank (0041000142825)	NULL	24500000	2020-04-03 00:00:00.000	Trangdtl	0
5	20200005	SKAI41030JP	Full phí	2020-04-06 00:00:00.000	Nhật 1 (Mai - 10160-68856611)	50000	10750000	2020-04-03 00:00:00.000	Ninhnht	0
6	20200006	SKAI31030JP	Full phí	2020-04-07 00:00:00.000	Vietcombank (0041000142825)	NULL	5400000	2020-04-04 00:00:00.000	Ninhnht	0
7	20200007	SKAI31030JP	Full phí	2020-04-08 00:00:00.000	Vietcombank (0041000142825)	NULL	7500000	2020-04-05 00:00:00.000	Trangdtl	0
8	20200008	SKAI11030VN	Full phí	2020-04-06 00:00:00.000	Vietinbank (103870317215)	NULL	5400000	2020-04-03 00:00:00.000	Trangdtl	0
9	20200009	CKAI121030VN	Full phí	2020-04-09 00:00:00.000	Vietcombank (0041000142825)	NULL	9500000	2020-04-06 00:00:00.000	Thaott	0
10	20200010	SKAI11030VN	Full phí	2020-04-09 00:00:00.000	Vietcombank (0041000142825)	NULL	5400000	2020-04-06 00:00:00.000	Ninhnht	0
11	20200011	SKAI21030VN	Full phí	2020-04-11 00:00:00.000	Nhật 1 (Mai - 10160-68856611)	25500	5432500	2020-04-09 00:00:00.000	Trangdtl	0
12	20200012	CKAI21030VN	Full phí	2020-04-11 00:00:00.000	Vietcombank (0041000142825)	NULL	9500000	2020-04-09 00:00:00.000	Trangdtl	0

Hình 4.32 Kết quả thu được tại bảng Fact_Sales

4.4.4. Data flow bảng QLHoc

Bảng FACT_QLHOC có được dữ liệu từ liệu từ việc bảng DIM_HOC merge join DIM_THANHTOAN sau đó thu được kết quả thu được merge join với bảng DIM_DANGKY thu được kết quả lọc những dữ liệu null thông qua data flow (Conditional Split) cuối cùng merge với bảng DIM_DATE.



Hình 4.33 Data flow bảng Fact_QLHoc

	ID_HV	ID_KH	ID_GV	Trangthailop	NgayBD	NgayKT_BL	Ngayhoc	Tongbobuoil	Sobuiconlai	TrangthaiTT	Ngaydk	Date_id
14	420246641	SKAI31030VN	GV239	Đang học	2022-08-23 00:00:00.000	NULL	NULL	32	27	Full phí	NULL	NULL
15	72019002	SKAI3N060VN	GV47	Kết thúc	2019-07-24 00:00:00.000	2019-11-11 00:00:00.000	T3-T5-T7	30	0	Đãt cọc	2019...	2019...
16	72019001	SKAI1N060VN	GV89	Kết thúc	2019-07-24 00:00:00.000	2019-11-11 00:00:00.000	T3-T5-T7	30	0	Đãt cọc	2019...	2019...
17	72019003	SKAI11030VN	GV104	Kết thúc	2019-07-27 00:00:00.000	2019-11-14 00:00:00.000	T3-T5-T7	30	0	Đãt cọc	2019...	2019...
18	72019004	SKAI31030VN	GV105	Kết thúc	2019-07-29 00:00:00.000	2019-11-16 00:00:00.000	T3-T5-T7	30	0	Đãt cọc	2019...	2019...
19	72019005	SKAI51030JP	GV01	Kết thúc	2019-07-30 00:00:00.000	2019-11-17 00:00:00.000	T3-T5-T7	30	0	Đãt cọc	2019...	2019...
20	72019006	SKAI2N060VN	GV106	Kết thúc	2019-07-31 00:00:00.000	2019-11-18 00:00:00.000	T3-T5-T7	30	0	Đãt cọc	2019...	2019...
21	72019007	SKAI41030JP	GV02	Kết thúc	2019-07-31 00:00:00.000	2019-11-18 00:00:00.000	T2-T4-T6	30	0	Đãt cọc	2019...	2019...
22	72019008	SKAI51030JP	GV03	Kết thúc	2019-08-03 00:00:00.000	2019-11-21 00:00:00.000	T3-T5-T7	30	0	Đãt cọc	2019...	2019...
23	72019009	SKAI11030VN	GV127	Kết thúc	2019-08-04 00:00:00.000	2019-11-22 00:00:00.000	T3-T5-T7	30	0	Đãt cọc	2019...	2019...
24	72019010	SKAI31030VN	GV129	Kết thúc	2019-08-05 00:00:00.000	2019-11-23 00:00:00.000	T3-T5-T7	30	0	Đãt cọc	2019...	2019...
25	72019011	SKAI31030VN	GV130	Kết thúc	2019-08-05 00:00:00.000	2019-11-23 00:00:00.000	T3-T5-T7	30	0	Đãt cọc	2019...	2019...
26	72019012	SKAI11030VN	GV125	Kết thúc	2019-08-07 00:00:00.000	2019-11-05 00:00:00.000	T3-T4-T5	30	0	Đãt cọc	2019...	2019...

Hình 4.34 Kết quả thu được tại bảng Fact_QLHoc

CHƯƠNG 5. CẬP NHẬT DỮ LIỆU BẢNG DIM

Kho dữ liệu lưu trữ dữ liệu lịch sử từ hệ thống xử lý giao dịch trực tuyến (Online transactional processing - OLTP). Khi dữ liệu mới được trích xuất vào kho dữ liệu từ hệ thống OLTP nguồn, một số bản ghi có thể thay đổi. Khi các thuộc tính của một bảng Dim nhất định thay đổi, đây được gọi là Dimension thay đổi chậm (Slowly Changing Dimension – SCD).

Có 3 kiểu xử lý thay đổi dữ liệu trong kho dữ liệu:

- Kiểu SCD1 (Type 1 Slowly Changing Dimension)
- Kiểu SCD2 (Type 2 Slowly Changing Dimension)
- Kiểu SCD3 (Type 3 Slowly Changing Dimension)

5.1. Ba kiểu cập nhật dữ liệu trên bảng Dimension

5.1.1. Kiểu SCD1 (Type 1 Slowly Changing Dimension)

Update, ghi đè dữ liệu vào bảng dimension. Kiểu xử lý này được sử dụng khi dữ liệu nguồn thay đổi dạng sửa sai hoặc khi phần cập nhật này không gây ảnh hưởng đến ý nghĩa của bảng fact. Mặc dù kiểu SCD1 không duy trì lịch sử, nhưng đây là cách đơn giản và nhanh nhất để tái dữ liệu bảng dimension.

Type 1 Slowly Changing Dimension							
Product Dim (Source)			Product Dim (Target)				
Product Name	Product ID	Product Descr		Product Name	SID	Source Product ID	Product Descr
10 inch box	010	10 inch glued box 10 inch pasted box	→	10 inch box	0001	010	10 inch pasted box
12 inch box	012	12 inch glued box	→	12 inch box	0002	012	12 inch glued box

Hình 5.1 Minh họa SCD kiểu 1 (Nguồn: Understanding Slowly Changing Dimensions, Oracle)

Hạn chế: Sau khi thực hiện có thể thấy rằng một số báo cáo phụ thuộc vào giá trị đó sẽ không trả lại thông tin giống như trước đây.

5.1.2. Kiểu SCD2 (Type 2 Slowly Changing Dimension)

Kiểu SCD2 (Type 2 Slowly Changing Dimension) cho phép theo dõi các thay đổi xảy ra trong bảng nguồn và liên kết chính xác bản ghi ở bảng đích. Khi kho dữ liệu nhận ra dữ liệu nguồn có sự thay đổi, thay vì ghi đè thì hệ thống cập nhật trạng thái bản ghi cũ và sinh thêm một bản ghi mới vào bảng đích. Bản ghi này được gán cho một khóa thay thế mới.

Kiểu SCD2 này thể hiện rõ nhất sự thay đổi của dữ liệu theo dòng thời gian vì mỗi sự thay đổi dù nhỏ nhất của thực thể trên dữ liệu nguồn đều được ghi nhận trong kho dữ liệu. Một vấn đề gặp phải là nếu số lượng bản ghi quá lớn sẽ khó khăn trong việc rà soát dữ liệu. Thay vì tìm bản ghi có khóa thay thế mới nhất (tốn tài nguyên, hiệu năng hệ thống) thì thường bổ sung trường STATUS để tiện truy vấn. Sau khi ghi nhận bản ghi mới thì cập nhật giá trị STATUS bản ghi cũ về trạng thái không sử dụng. Một số thông tin bổ sung thường được lưu lại bao gồm:

- Ngày bắt đầu có hiệu lực (START_TIME)
- Ngày hết hiệu lực (sau khi hết hiệu lực mới được cập nhật thêm) (END_TIME)
- Lý do cập nhật

Product Dim (Source)			Product Dim (Target)					
Product Name	Product ID	Product Descr	SID	Source Product ID	Product Name	Product Descr	EFF_START_DT	EFF_END_DT
12 inch box	012	12 inch glued box	0001	012	12 inch box	12 inch glued box	Jan-01-1753	Dec-31-9999
10 inch box	010	10 inch glued box 10 inch pasted box	0002	010	10 inch box	10 inch glued box	Jan-01-1753	May-12-06
			0003	010	10 inch box	10 inch pasted box	May-12-06	Dec-31-9999

Hình 5.2 Minh họa SCD kiểu 2 (Nguồn: Understanding Slowly Changing Dimensions, Oracle)

Hạn chế: cần phải tổng quát hóa khóa bảng Dim và sự phát triển của chính kích thước bảng. Bảng Dim có thể trở nên khá lớn trong trường hợp có một số thay đổi đối với các thuộc tính thứ nguyên được theo dõi.

5.1.3. Kiểu SCD3 (Type 3 Slowly Changing Dimension)

Kiểu SCD3 (Type 3 Slowly Changing Dimension) tạo ra một cột giá trị hiện tại mới trong bản ghi hiện có nhưng vẫn giữ lại cột ban đầu. Cột giá trị hiện tại mới chứa dữ liệu bản ghi mới đến từ hệ thống OLTP. Kiểu SCD3 này được sử dụng khi thay đổi trong bản ghi của bảng Dim phải được theo dõi nhưng giá trị cũ phải được giữ lại như một phần của bản ghi, thường là để báo cáo.

Hạn chế: chỉ xử lý hai thay đổi gần đây nhất. Nếu nhiều thay đổi diễn ra và tất cả chúng phải được theo dõi thì nên sử dụng kiểu SCD2. Nhìn chung kiểu SCD3 ít được sử dụng.

5.2. Thực hành cập nhật dữ liệu trên bảng Dimension

Để phù hợp với quy trình nghiệp vụ và thiết kế hệ của các báo cáo thống kê, nhóm chọn phương án kết hợp kiểu SCD1 và kiểu SCD2. Khi đó các thuộc tính được đánh ưu tiên, nếu thuộc tính thay đổi thuộc độ ưu tiên thấp thì dùng kiểu SCD1, ghi đè giá trị lên tất cả các bản ghi lịch sử từ trước đến nay. Nếu thuộc tính thay đổi thuộc độ ưu tiên cao hơn thì dùng kiểu SCD2, vô hiệu hóa bản ghi cũ và sinh ra bản ghi mới.

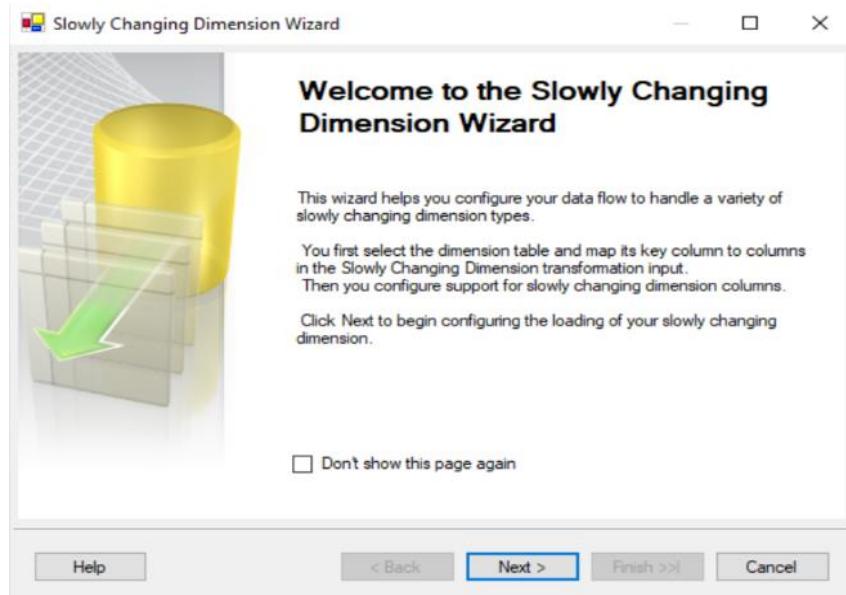
Cụ thể nhóm thực hiện cập nhật dữ liệu kiểu SCD1 kết hợp kiểu SCD2 trên bảng Dim_HocVien với các thuộc tính lựa chọn như sau:

- “Email”: độ ưu tiên thấp, cập nhật theo kiểu SCD1.
- “TrinhDo”: độ ưu tiên cao hơn, cập nhật theo kiểu SCD2.

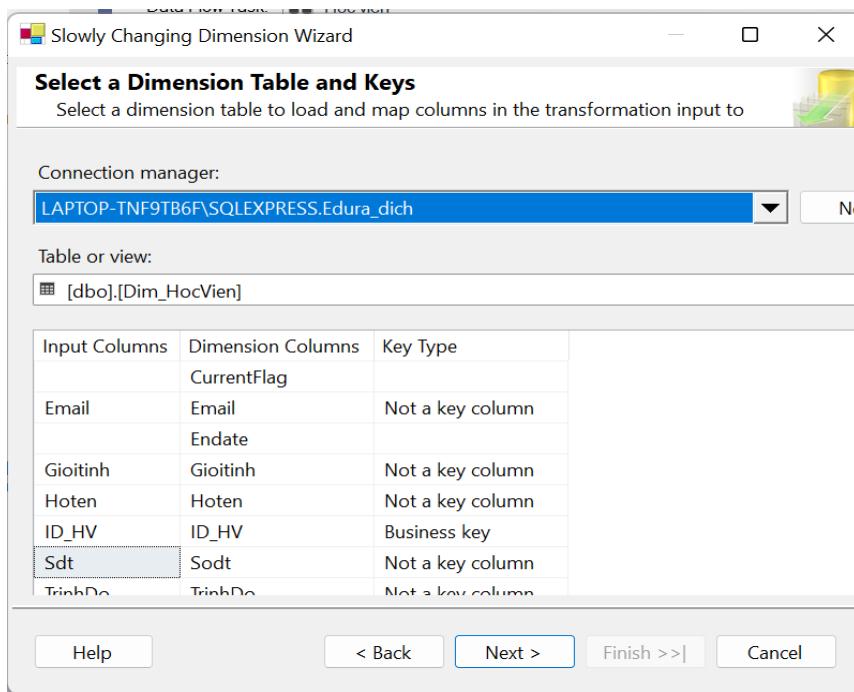
Quy trình tiến hành cập nhật dữ liệu trên bảng Dim_HocVien được thực hiện qua 3 bước: Set up công cụ SCD, đồ lại dữ liệu và thực hiện cập nhật dữ liệu. Cụ thể:

Bước 1: Set up công cụ Slowly Changing Dimension

- Mở giao diện Slowly Changing Dimension

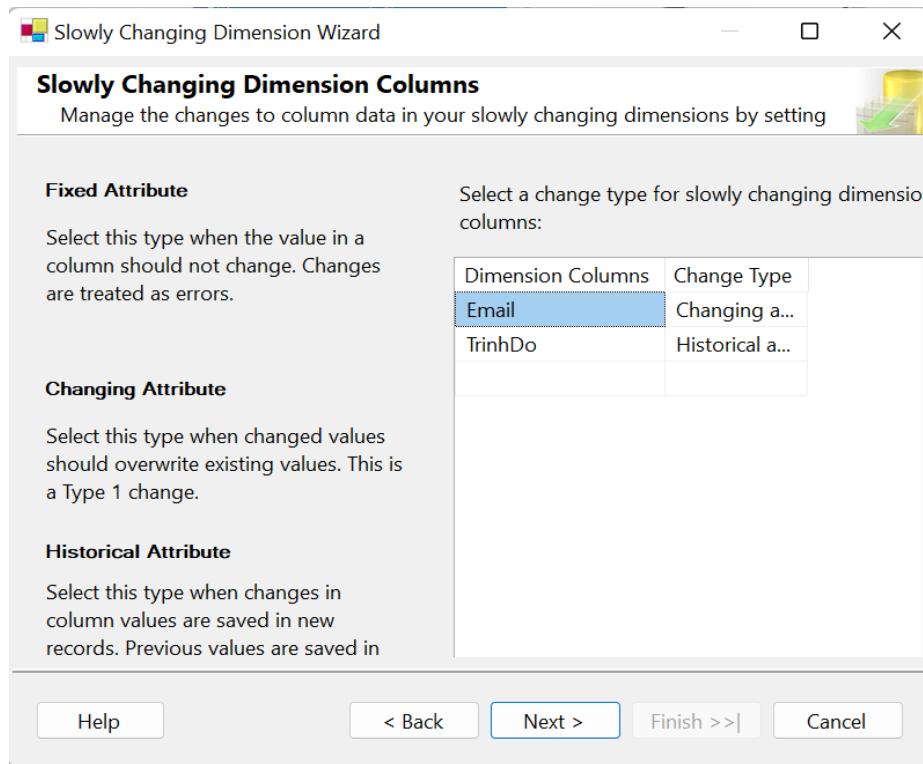


- Chọn bảng Dim_HocVien, chọn Key Type của ID_HV là Business key.

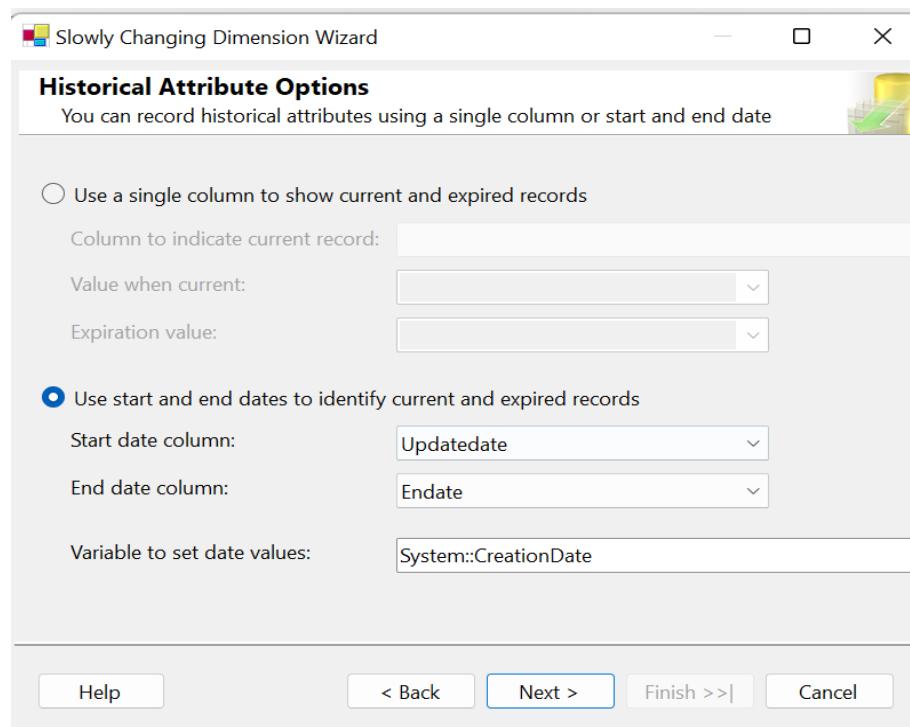


- Chọn thuộc tính và kiểu cập nhật tương ứng

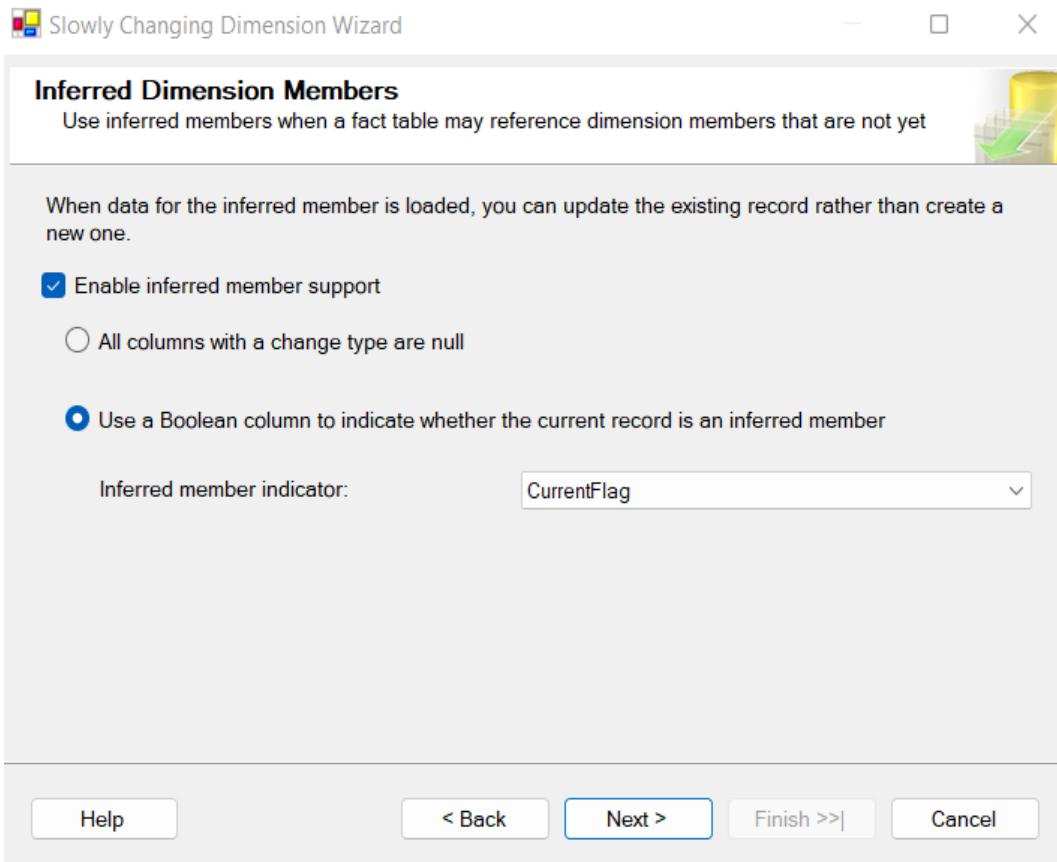
Dimension Columns	Change type
Email	Changing attribute
TrinhDo	Historical attribute



- Trường *Updatedate* và *Endate* phục vụ việc lưu lại lịch sử thay đổi trường “TrinhDo”.



- Trường CurrentFlag được sử dụng để đánh dấu dữ liệu mới hay cũ:



- Nhấn Finish

Bước 2: Chạy Data flow đồ dữ liệu lại vào Dim_HocVien

- Code Trigger cập nhật trường CurrentFlag

```
create trigger tg_trinhdo on Dim_HocVien
for insert, update as
begin
    update Dim_HocVien
    set CurrentFlag= 1 from inserted a, Dim_HocVien b
    where b.Endate is null and a.ID_HV=b.ID_HV

    update Dim_HocVien
    set CurrentFlag= 0 from inserted a, Dim_HocVien b
    where b.Endate < GETDATE() and a.ID_HV=b.ID_HV
end
GO
```

- Dữ liệu bảng Dim_HocVien

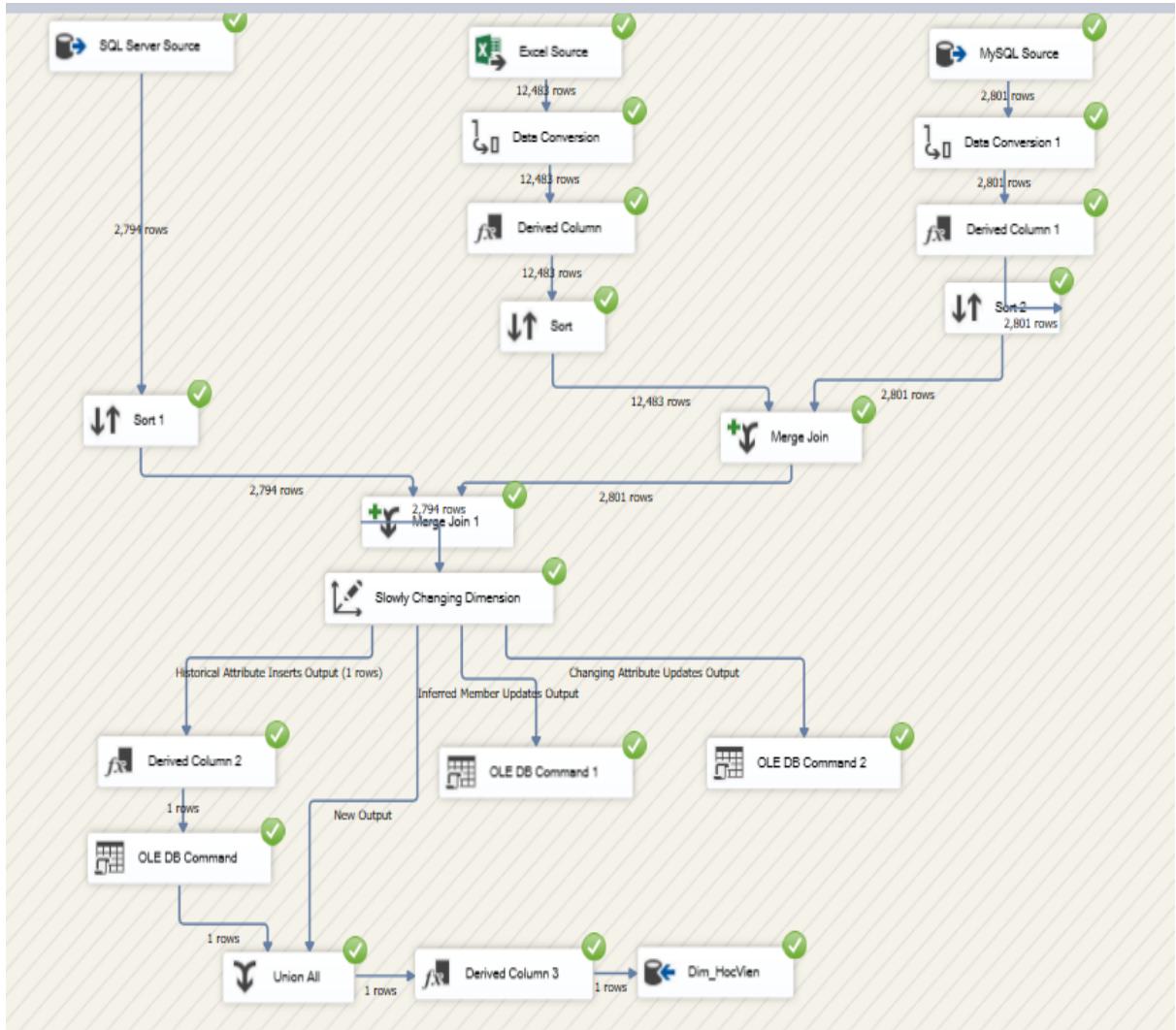
ID_HV	Hoten	Gioiti...	Sodt	Email	TrinhDo	Updatedate	End...	CurrentFlag
42020057	Nguyễn Thái Đan Tâm	Nữ	981101197	nguyenthaidantam@gmail.com	Kaiwa 2	2022-09-18	NULL	NULL
42020058	Xuân Bảo	Nam	989761637	xuanbao07041987@gmail.com	Kaiwa 1	2022-09-18	NULL	NULL
42020059	Nguyễn Ngọc Hồng ...	Nữ	937984818	duyennguyen1031994@gmail.com	Kaiwa 4	2022-09-18	NULL	NULL
42020060	Nguyễn Đăng Hiệu	Nam	975313642	danghieu3642@gmail.com	Kaiwa 2	2022-09-18	NULL	NULL
42020061	Dương Thị Bích Trâm	Nữ	908241995	bichtram199524@gmail.com	Kaiwa 3	2022-09-18	NULL	NULL
42020062	Nguyễn Thị Phương ...	Nữ	334331763	phuongtrinh.jps@gmail.com	Kaiwa 4	2022-09-18	NULL	NULL
42020063	Đặng Nguyễn Hồng ...	Nữ	935031188	dangphuc.mitsubishi@gmail.com	Kaiwa 3	2022-09-18	NULL	NULL
42020064	Nguyễn Thiện Nhân	Nam	909024896	nhannt0824@gmail.com	Kaiwa 2	2022-09-18	NULL	NULL
42020065	Nguyễn Bình	Nam	1648844588	nguyencongbinh555@gmail.com	Kaiwa 3	2022-09-18	NULL	NULL
42020066	Mai Huỳnh Thảo Nhi	Nữ	8090289727	mhtn.mikan@gmail.com	Kaiwa 3	2022-09-18	NULL	NULL
42020067	Nguyễn Văn Hùng	Nam	7038202234	Rogerhungnguyen@gmail.com	Kaiwa 4	2022-09-18	NULL	NULL
42020068	Phương Trúc Huyền	Nữ	8030846886	truchuyenphuong@gmail.com	Kaiwa 3	2022-09-18	NULL	NULL
42020069	Nguyễn Thị Bình	Nữ	963709278	Sn01041994@gmail.com	Kaiwa 3	2022-09-18	NULL	NULL
42020070	Phạm Hoàng My	Nữ	979632735	hoangmypham23591@gmail.com	Kaiwa 2	2022-09-18	NULL	NULL
42020071	Lê Thị Ngọc Anh	Nữ	8081830710	ngocanhaoyama@gmail.com	Kaiwa 3	2022-09-18	NULL	NULL
42020072	Trần Định Nam Kha	Nam	9063691312	trandinhnamkhabmt@gmail.com	Kaiwa 3	2022-09-18	NULL	NULL
42020073	Hoàng Thị Thùy Trang	Nữ	784956057	hoangthuytrang28@gmail.com	Kaiwa 3	2022-09-18	NULL	NULL
42020074	Lương Thị Hằng	Nữ	813771560	gianhubg@gmail.com	Kaiwa 1	2022-09-18	NULL	NULL
42020075	Lê Thuý Kiều	Nữ	855663747	Thuykieu240901@gmail.com	Kaiwa 1	2022-09-18	NULL	NULL
42020076	Nguyễn Thị Kim Yến	Nữ	706318448	yenxinhdep571999@gmail.com	Kaiwa 2	2022-09-18	NULL	NULL
42020077	Sơn Thành Đạt	Nam	NULL	sonthanhdat01@gmail.com	Kaiwa 3	2022-09-18	NULL	NULL
42020078	Lại Thị Miền	Nữ	369805049	laithimien.tcnh@gmail.com	Kaiwa 1	2022-09-18	NULL	NULL
42020079	Thu Hoài	Nữ	375497674	hoaiqt96@gmail.com	Kaiwa 3	2022-09-18	NULL	NULL
42020080	Nguyễn Tử Ân	Nữ	8091776526	kb1909jp.com@icloud.com	Kaiwa 1	2022-09-18	NULL	NULL
42021429	Ngoài Nguyễn Thái Đạt	Nam	8071425569	nauvienthaidat1986@gmail.com	Kaiwa 3	2022-09-18	NULL	NULL

Bước 3: Update dữ liệu bảng HocVien\$ (Nguồn 1 – SQL Server Source) và kiểm tra kết quả.

- Update TrinhDo của học viên có ID=20200007 lên Kaiwa 4

```
update HocVien$
set TrinhDo = 'Kaiwa 4'
where ID_HV = 20200007;
```

- Cập nhật thành công trên SSDT



- Dữ liệu bảng Dim_HocVien sau khi cập nhật trình độ theo kiểu SCD2

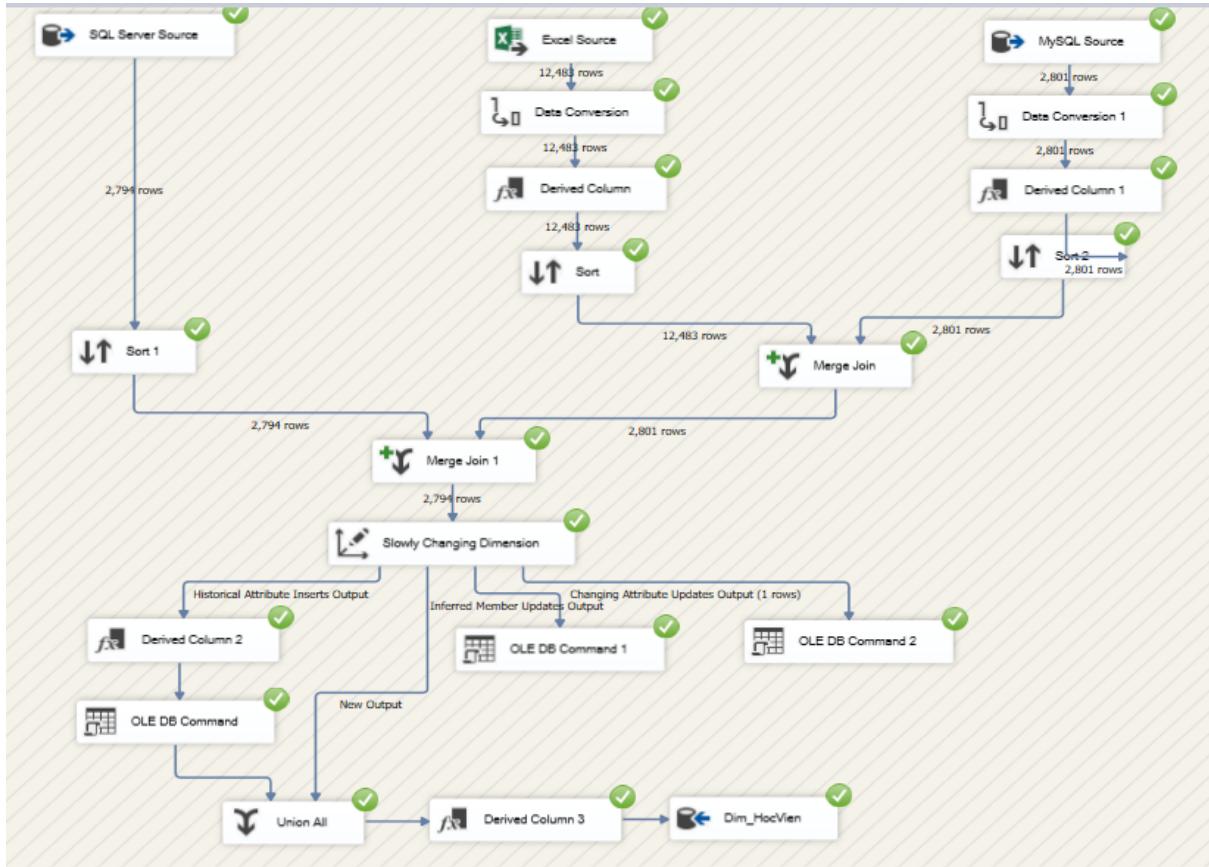
	ID_HV	Hoten	Gioitinh	Sodt	Email	TrinhDo	Updatedate	Endate	CurrentFlag
1	20200007	Mai Thiên Vũ	Nam	902380306	thienvumai@gmail.com	Kaiwa 3	2022-09-18	2022-10-11	0
2	20200007	Mai Thiên Vũ	Nam	NULL	thienvumai@gmail.com	Kaiwa 4	2022-10-11	NULL	1

- Update Email của học viên có ID=20200008 thành “tripv2001@gmail.com”

```

update HocVien$
set Email = 'tripv2001@gmail.com'
where ID_HV = 20200008;
  
```

- Cập nhật thành công trên SSDT



- Dữ liệu bảng trước khi cập nhật Email

	ID_HV	Hoten	Gioitinh	Sodt	Email	TrinhDo	Updatedate	Endate	CurrentFlag
1	20200008	Phan Văn Trí	Nam	972561243	phanvantrixd4a2@gmail.com	Kaiwa 1	2022-09-18	NULL	1

- Dữ liệu sau khi cập nhật Email kiểu SCD1

	ID_HV	Hoten	Gioitinh	Sodt	Email	TrinhDo	Updatedate	Endate	CurrentFlag
1	20200008	Phan Văn Trí	Nam	972561243	tripv2001@gmail.com	Kaiwa 1	2022-09-18	NULL	1

Như vậy, chúng ta dễ dàng cập nhật lại Email của học viên nếu như muốn sửa đổi trong trường hợp nhập sai thông tin hoặc có nhu cầu thay đổi tại nguồn. Với các thông tin cần lưu lại lịch sử, cụ thể là thuộc tính trình độ của học viên giúp trung tâm dễ dàng nắm được sự tiến bộ của học viên, giúp trung tâm dễ dàng đánh giá và cập nhật học liệu và học viên có thể đạt được mục tiêu học tập nhanh nhất.

CHƯƠNG 6. XÂY DỰNG VÀ TRIỂN KHAI DATA LAKE TRÊN AMAZON WEB SERVICE (AWS)

6.1. Tìm hiểu giải pháp xây dựng Data Lake thu thập dữ liệu phi cấu trúc cho doanh nghiệp EDTE

Hiện nay, với các dữ liệu sẵn có của công ty được mỗi phòng ban khai thác và quản lý những trường dữ liệu đặc thù của mình mà bỏ qua khá nhiều nguồn dữ liệu khác được thu thập như: dữ liệu truyền thông trên mạng xã hội, các tài liệu khóa học có như các video, video lịch sử các lớp học trực tuyến, các đoạn tin nhắn/ cuộc trò chuyện được trao đổi giữa các học viên với giảng viên... bởi hầu hết những dữ liệu này là các dữ liệu phi cấu trúc khó tìm kiếm, không được lưu trữ trong các cơ sở dữ liệu hay kho dữ liệu.

Vậy tại sao phải cần lưu trữ các dữ liệu phi cấu trúc? Ví dụ từ những dữ liệu được định dạng bằng âm thanh khi nhân viên sale gọi điện để tư vấn khóa học cho học viên có nhu cầu thì từ các file dữ liệu đó chúng ta có thể nhận dạng giọng nói để xác định học viên tiềm năng và thu thập thông tin về các truy vấn và cảm xúc của họ để có thể xây dựng hình thức marketing phù hợp với nhu cầu của học viên. Hay từ các dữ liệu về đoạn chat chúng ta có thể dễ dàng thống kê lại các chủ đề, thuộc tính, mối quan tâm thường thấy, các câu hỏi hay được đề cập đến để có thể lên kế hoạch xây dựng chatbots hỗ trợ trong tư vấn từ đó làm giảm thời gian tư vấn cũng như tối ưu hóa được quá trình chăm sóc khách hàng như vậy sẽ làm cho doanh nghiệp tiết kiệm chi phí, thời gian, tránh xử lý các công việc lặp đi lặp lại nhiều lần. Vậy nên để có thể lưu trữ quản lý và sử dụng cho các công việc phân tích thì chúng ta có thể xây dựng Data Lake để có thể lưu trữ chúng một cách dễ dàng.

Công cụ có thể xây dựng Data Lake khá phổ biến như: Azure Data Lake Storage, Amazon S3, Qubole, Snowflake... Nhưng với giải pháp sử dụng dịch vụ để lưu trữ dữ liệu phi cấu trúc bao gồm hình ảnh, video, âm thanh, tin nhắn... thì sử dụng dịch vụ Amazon S3 là phù hợp nhất. Vì Data Lake trên AWS tận dụng tính bảo mật, độ bền và khả năng mở rộng của Amazon S3 để quản lý danh mục liên tục gồm các tập dữ liệu tổ chức và Amazon DynamoDB để quản lý siêu dữ liệu tương ứng. Sau khi tập dữ liệu

được lập danh mục, các thuộc tính và thẻ mô tả của nó sẽ có sẵn để tìm kiếm. Người dùng có thể tìm kiếm và duyệt qua các tập dữ liệu có sẵn trong bảng điều khiển, đồng thời tạo danh sách dữ liệu mà họ yêu cầu quyền truy cập. Nó theo dõi các tập dữ liệu mà người dùng chọn và tạo tệp kê khai với các liên kết truy cập an toàn đến nội dung mong muốn khi người dùng kiểm tra.

6.2. Đề xuất doanh nghiệp EDTE nên nâng cấp hệ thống hiện tại lên Cloud và ứng dụng giải pháp Data Lake trên AWS

6.2.1. Lý do doanh nghiệp cần lên kế hoạch nâng cấp hệ thống hiện tại lên Cloud và ứng dụng Data Lake

Để xây dựng một chiến lược kinh doanh và hiệu quả thì việc tiến hành xây dựng các kế hoạch dài hạn là điều cần thiết cho doanh nghiệp EDTE vì vậy việc cân nhắc nâng cấp hệ thống hiện tại lên Cloud/ứng dụng Data Lake là điều cần thiết trong kế hoạch này. Vậy tại sao doanh nghiệp cần phải nâng cấp hệ thống hiện tại của mình? Dưới đây là một số phân tích trong kế hoạch tương lai mà doanh nghiệp hướng đến:

Mở rộng quy mô đào tạo từ online sang kết hợp với hệ đào tạo offline

Hiện nay, học viên từ EDTE chủ yếu đến từ Việt Nam và một số ít là các du học sinh đang học tập, sinh sống và làm việc tại Nhật Bản. Và mô hình hoạt động của doanh nghiệp đang xây dựng đều là hình thức học online vậy nên trong tương lai doanh nghiệp mong muốn mở rộng thêm các loại hình dịch vụ và phát triển thêm hệ thống học tập offline tại các địa điểm trung tâm để thu hút nhiều nguồn khách hàng hơn. Như vậy sẽ khiến doanh nghiệp có nguồn học viên phong phú và hoạt động marketing sẽ đạt được hiệu quả cao hơn thay vì chỉ hoạt động marketing online như hiện nay.

Phân tích phản hồi từ học viên để từ đó thay đổi các chính sách cũng như nội dung chương trình đào tạo

Với các dữ liệu phản hồi từ học viên thì EDTE đang quản lý rời rạc chưa có hệ thống riêng biệt để quản lý vì vậy các vấn đề được học viên phản ánh rất dễ bị bỏ sót từ đó khiến cho quy trình chăm sóc học viên bị ảnh hưởng, nội dung chương trình đào tạo sẽ không được cập nhật liên tục gây nhảm chán cho học viên. Chính vì vậy, doanh nghiệp

cần tổng hợp chúng để lưu trữ ở một nơi thống nhất để tiện phân tích tránh được bỏ sót thông tin đáng giá từ đó xác định được vấn đề cần phải giải quyết trong quy trình đào tạo của mình

Thu thập nội dung chat để xây dựng mô hình Chatbot

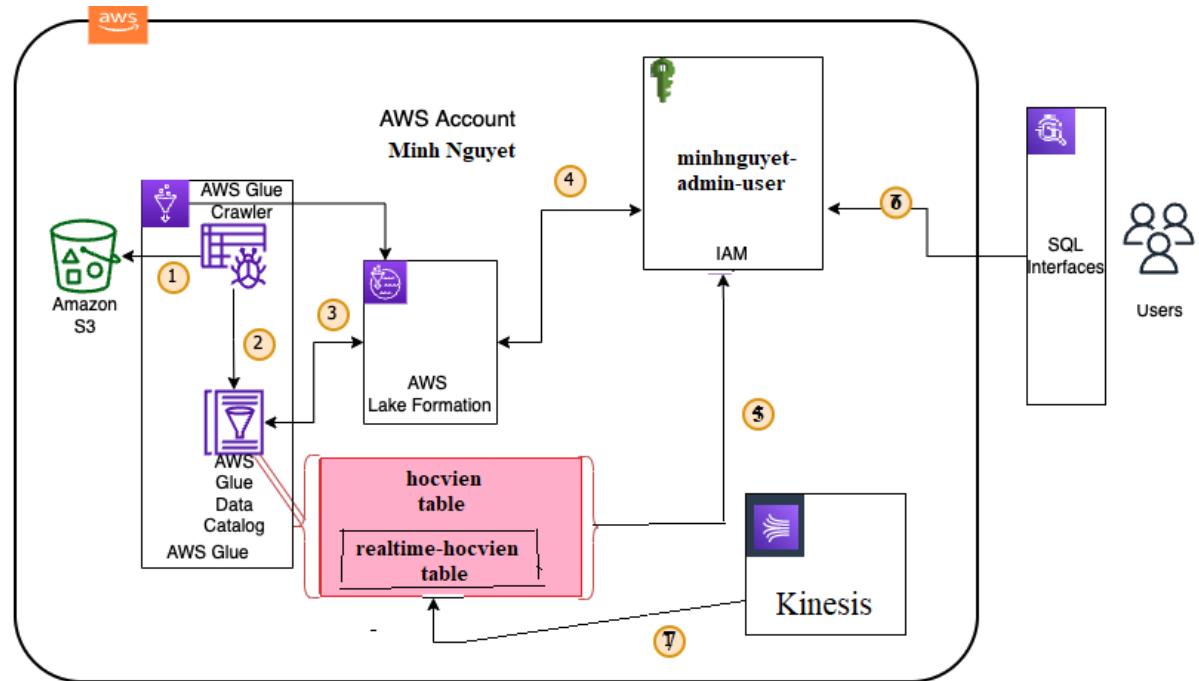
Với các nguồn thông tin từ các đoạn chat thì tỷ lệ lặp lại chủ đề, câu hỏi là điều hết sức bình thường. Và đội ngũ nhân viên tư vấn hỗ trợ sẽ phải trả lời lại những câu hỏi, chủ đề được lặp đi lặp lại là điều gây phiền não vì nó sẽ làm mất thời gian và chi phí để hoạt động. Vậy nên việc thu thập lại các vấn đề thường gặp để có thể xây dựng các kịch bản chatbot sẽ giúp doanh nghiệp giảm thiểu được chi phí, thời gian, tránh được các vấn đề phải xử lý lại nhiều lần.

6.2.2. Lợi ích doanh nghiệp sẽ nhận được nếu nâng cấp hệ thống hiện tại lên Cloud và ứng dụng Data Lake

Khi doanh nghiệp triển khai sử dụng Cloud/ Data Lake thì họ sẽ nhận được khá nhiều lợi ích từ việc này như: **Khả năng sử dụng và khả năng tiếp cận sẽ dễ dàng hơn** vì cơ bản các dịch vụ này đều tương tác thông qua giao diện vì vậy không đòi hỏi quá nhiều về kiến thức công nghệ. Ngoài ra khi quản lý trên đây tính bảo mật sẽ được siết chặt vì cơ bản trên đây hầu hết sẽ lưu trữ các dữ liệu dự phòng vì vậy ngay cả khi một trong các trung tâm dữ liệu bị sập, hỏng hóc, mất mát thì dữ liệu của doanh nghiệp sẽ được an toàn và được giám sát và một dịch vụ lưu trữ Cloud nào cũng được hình thành từ hàng nghìn trung tâm dữ liệu. Ngoài ra, việc này cũng giúp chúng ta tiết kiệm chi phí, chia sẻ tệp thuận tiện hay tự động hóa một số quy trình riêng...

6.3. Xây dựng và triển khai Data Lake trên Amazon Web Services

6.3.1. Các dịch vụ được sử dụng trên AWS để xây dựng Data Lake



Hình 6.1 Sơ đồ minh họa kiến trúc hồ dữ liệu sử dụng trên AWS

Giải thích các ký hiệu trên sơ đồ:

1. Dữ liệu được lưu trữ trong hồ dữ liệu Amazon S3 được thu thập thông tin bằng trình thu thập dữ liệu AWS Glue.
2. Trình thu thập thông tin cung cấp siêu dữ liệu của dữ liệu trên Amazon S3 và lưu trữ nó dưới dạng cơ sở dữ liệu và bảng trong Danh mục dữ liệu AWS Glue.
3. Đăng ký nhóm Amazon S3 làm vị trí hồ dữ liệu với Lake Formation. Nó được tích hợp nguyên bản với Data Catalog.
4. Sử dụng Lake Formation để cấp quyền ở cấp cơ sở dữ liệu, bảng và cột cho các vai trò AWS Identity and Access Management (IAM) đã xác định.
5. Cung cấp quyền truy cập cho các nhóm sales vào các lược đồ bên ngoài tương ứng của họ và liên kết các vai trò IAM thích hợp sẽ được đảm nhận. Vai trò quản trị viên và nhóm quản trị viên bị giới hạn đối với công việc quản trị.

6. Người dùng sales hiện có thể đảm nhận vai trò IAM tương ứng của họ và truy vấn dữ liệu bằng cách sử dụng trình chỉnh sửa truy vấn SQL trong Athena
7. Sử dụng dịch vụ Kinesis để nhập dữ liệu truyền trực tiếp theo thời gian thực vào hồ dữ liệu.

Ngoài ra, nhóm cũng sử dụng AWS Glue ETL để chuyển đổi dữ liệu json dòng thời gian thực của chúng tôi thành Parquet thông qua Glue Job. Sau đó, nhóm sử dụng AWS Glue Crawler để tạo một bảng mới trên dữ liệu Parquet này trong cơ sở dữ liệu mới. Quyền truy cập vào dữ liệu được kiểm soát bởi AWS Lake Formation.

6.3.2. Thực hành xây dựng Data Lake trên AWS

Bước 1: Add user

- Set user details

Add user

Set user details

You can add multiple users at once with the same access type and permissions. [Learn more](#)

User name*

[Add another user](#)

Select AWS access type

Select how these users will primarily access AWS. If you choose only programmatic access, it does NOT prevent users from accessing the console using an assumed role. Access keys and autogenerated passwords are provided in the last step. [Learn more](#)

Select AWS credential type* **Access key - Programmatic access**
Enables an **access key ID** and **secret access key** for the AWS API, CLI, SDK, and other development tools.

Password - AWS Management Console access
Enables a **password** that allows users to sign-in to the AWS Management Console.

Console password* Autogenerated password Custom password

 Show password

Require password reset User must create a new password at next sign-in
Users automatically get the [IAMUserChangePassword](#) policy to allow them to change their own password.

* Required

Cancel [Next: Permissions](#)

- Set permissions

Add user

1 2 3 4 5

Review

Review your choices. After you create the user, you can view and download the autogenerated password and access key.

User details

User name	minhnguyet-admin-user
AWS access type	AWS Management Console access - with a password
Console password type	Custom
Require password reset	No
Permissions boundary	Permissions boundary is not set

Permissions summary

The following policies will be attached to the user shown above.

Type	Name
Managed policy	AdministratorAccess
Managed policy	AWSLakeFormationDataAdmin

Tags

No tags were added.

Cancel Previous Create user

⇒ Kết quả tạo User thành công

IAM > Users

Users (2) Info

An IAM user is an identity with long-term credentials that is used to interact with AWS in an account.

<input type="checkbox"/>	User name	Groups	Last activity	MFA	Password a...	Active key age
<input type="checkbox"/>	minhnguyet-admin-user	None	Never	None	<input checked="" type="checkbox"/> Now	-

Bước 2: AWS Lake Formation

- Thêm Administrative roles and tasks

The screenshot shows the 'Data lake administrators (0/2)' section. It includes a search bar labeled 'Find administrators' and a table with two columns: 'Name' and 'Type'. A single row is listed: 'minhnguyet-admin-user' under 'Name' and 'IAM user' under 'Type'.

Name	Type
minhnguyet-admin-user	IAM user

- Đăng nhập AWS với user vừa tạo

Sign in as IAM user

Account ID (12 digits) or account alias

IAM user name

Password

Remember this account

Sign in

- Grant Administrative roled and tasks

Grant permissions

Choose the access permissions to grant.

IAM users and roles
Add one or more IAM users or roles.

Choose IAM principals to add

minhnguyet-admin-user X
User

SAML and Amazon QuickSight users and groups
Enter a SAML user or group ARN or Amazon QuickSight ARN. Press Enter to add additional ARNs.

Ex: arn:aws:iam::<AccountId>:saml-provider/<SamlProviderName>:user/<UserName>

Catalog permissions
Choose the specific access permissions to grant.

Create database

Grantable permissions
Choose the permissions that may be granted to others.

Create database

Cancel **Grant**

⇒ Kết quả thành công

Database creators (0/3)

Choose IAM principals permitted to create databases in your AWS Glue Data Catalog.

Principal	Principal type	Permissions	Grantable
saleuser	IAM user	Create database	-
minhnguyet-admin-user	IAM user	Create database	Create database

Bước 3: Tạo Bucket và upload dữ liệu

- Tạo bucket

Create bucket Info

Buckets are containers for data stored in S3. [Learn more](#)

General configuration

Bucket name

Bucket name must be globally unique and must not contain spaces or uppercase letters. [See rules for bucket naming](#)

AWS Region

Asia Pacific (Singapore) ap-southeast-1

Copy settings from existing bucket - *optional*
Only the bucket settings in the following configuration are copied.

Choose bucket

⇒ Kết quả tạo bucket thành công

Buckets (2)				<small>Info</small>			
				Buckets are containers for data stored in S3. Learn more			
	Name	AWS Region	Access				
	minhnguyet-bucket	Asia Pacific (Singapore) ap-southeast-1	Bucket and objects not public				

- Upload dữ liệu

Amazon S3 > Buckets > minhnguyet-bucket

minhnguyet-bucket [Info](#)

[Objects](#) [Properties](#) [Permissions](#) [Metrics](#) [Management](#) [Access Points](#)

Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

[Delete](#) [Actions ▾](#) [Create folder](#)

[Find objects by prefix](#) [1](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	AthenaResults/	Folder	-	-	-
<input type="checkbox"/>	data/	Folder	-	-	-

Amazon S3 > Buckets > minhnguyet-bucket > data/ > sales/

sales/

[Objects](#) [Properties](#)

Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

[Delete](#) [Actions ▾](#) [Create folder](#)

[Find objects by prefix](#) [1](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	hocvien/	Folder	-	-	-
<input type="checkbox"/>	real-time-hocvien/	Folder	-	-	-

⇒ Kết quả upload thành công

Amazon S3 > Buckets > minhnguyet-bucket > data/ > sales/ > hocvien/

hocvien/

Objects (1)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Actions Copy S3 URI Copy URL Download Open Delete Create folder Upload

Find objects by prefix

Name	Type	Last modified	Size	Storage class
sales.csv	csv	October 13, 2022, 11:01:13 (UTC+07:00)	213.1 KB	Standard

- Đăng ký vị trí

Register location

Amazon S3 location

Register an Amazon S3 path as the storage location for your data lake.

Amazon S3 path

Choose an Amazon S3 path for your data lake.

s3://minhnguyet-bucket

Review location permissions - strongly recommended

Registering the selected location may result in your users gaining access to data already at that location. Before registering a location, we recommend that you review existing location permissions on resources in that location.

IAM role

To add or update data, Lake Formation needs read/write access to the chosen Amazon S3 path. Choose a role that you know has permission to do this, or choose the **AWSServiceRoleForLakeFormationDataAccess** service-linked role. When you register the first Amazon S3 path, the service-linked role and a new inline policy are created on your behalf. Lake Formation adds the first path to the inline policy and attaches it to the service-linked role. When you register subsequent paths, Lake Formation adds the path to the existing policy.

AWSServiceRoleForLakeFormationDataAccess

⚠️ Do not select the service linked role if you plan to use EMR.

Bước 4: Tạo database

Create database

Database details
Create a database in the AWS Glue Data Catalog.

Database
Create a database in my account.

Resource link
Create a resource link to a shared database.

Name

Location - optional
Choose an Amazon S3 path for this database, which eliminates the need to grant data location permissions on catalog table paths that are this location's children

Description - optional

Descriptions can be up to 2048 characters long.

Default permissions for newly created tables
This setting maintains existing AWS Glue Data Catalog behavior. You can still set individual permissions, which will take effect when you revoke the Super permission from IAMAllowedPrincipals. See [Changing Default Settings for Your Data Lake](#).

Use only IAM access control for new tables in this database

[Cancel](#) [Create database](#)

⇒ Kết quả đạt được

AWS Lake Formation > Databases							
Databases (0/2)							
<input type="button" value="C"/> Actions							
<input type="text"/> Find databases							
	Name	▲	Owner account ID	▼	Shared resource	▼	Shared resource owner
<input type="radio"/>	minhnguyetdb		734081627824	-	-	-	-

Bước 5: Tạo Crawler

Add information about your crawler

Crawler name

SalesCrawlerHocVien

- ▶ Tags, description, security configuration, and classifiers (optional)

Next

Specify crawler source type

Choose Existing catalog tables to specify catalog tables as the crawler source. The selected tables specify the data stores to crawl. This option doesn't support JDBC data stores.

Crawler source type

- Data stores
- Existing catalog tables

Repeat crawls of S3 data stores

- Crawl all folders

Crawl all folders again with every subsequent crawl.
- Crawl new folders only

Only Amazon S3 folders that were added since the last crawl will be crawled. If the schemas are compatible, new partitions will be added to existing tables.
- Crawl changed folders identified by Amazon S3 Event Notifications

Rely on Amazon S3 events to control what folders to crawl.

Back

Next

Add a data store

Choose a data store

S3

Connection

Select a connection

Optional include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any future S3 targets will also use the same connection (or none, if left blank).

[Add connection](#)

Crawl data in

Specified path in my account
 Specified path in another account

Include path

s3://minhnguyet-bucket/data/sales/hocvien

All folders and files contained in the include path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.

Sample size (optional)

Enter a number between 1 and 249

This field sets the number of files in each leaf folder to be crawled. If not set, all the files are crawled.

► Exclude patterns (optional)

[Back](#) [Next](#)

Add another data store

Yes
 No

[Back](#) [Next](#)

Choose an IAM role

The IAM role allows the crawler to run and access your Amazon S3 data stores. [Learn more](#)

- Update a policy in an IAM role
- Choose an existing IAM role
- Create an IAM role

IAM role

AWSGlueServiceRole-

To create an IAM role, you must have **CreateRole**, **CreatePolicy**, and **AttachRolePolicy** permissions.

Create an IAM role named "**AWSGlueServiceRole**-rolename" and attach the AWS managed policy, **AWSGlueServiceRole**, plus an inline policy that allows read access to:

- s3://minhnguyet-bucket/data/sales/hocvien

You can also create an IAM role on the [IAM console](#).

[Back](#)

[Next](#)

Create a schedule for this crawler

Frequency

Run on demand



[Back](#)

[Next](#)

Configure the crawler's output

Database i

minhnguyetdb ▼

Add database

Prefix added to tables (optional) i

Type a prefix added to table names

Table threshold (optional)

Enter a number greater than 0

This field sets the maximum number of tables the crawler is allowed to generate. In the event that this number is surpassed, the crawl will fail with an error. If not set, the crawler will automatically generate the number of tables depending on the data schema.

- ▶ Grouping behavior for S3 data (optional)
- ▶ Configuration options (optional)

Back Next

Crawler info

Name SalesCrawlerHocVien
Tags -

Data stores

Data store S3
Include path s3://minhnguyet-bucket/data/sales/hocvien
Connection
Exclude patterns

IAM role

IAM role arn:aws:iam::734081627824:role/service-role/AWSGlueServiceRole-SalesHocVienCrawler

Schedule

Schedule Run on demand

Output

Database minhnguyetdb
Prefix added to tables (optional)
Table threshold (optional)
Create a single schema for each S3 path false
Table level (optional)
▶ Configuration options

[Back](#) [Finish](#)

Bước 6: Tạo Grant Data permissions

AWS Lake Formation > Grant permissions

Grant data permissions

Principals

- IAM users and roles
Users or roles from this AWS account.
- SAML users and groups
SAML users and group or QuickSight ARNs.
- External accounts
AWS accounts or AWS organizations outside of this account.

IAM users and roles
Add one or more IAM users or roles.

Choose IAM principals to add ▾

AWSGlueServiceRole-SalesHocVienCrawler X
Role

LF-Tags or catalog resources

- Resources matched by LF-Tags (recommended)
Manage permissions indirectly for resources or data matched by a specific set of LF-Tags.
- Named data catalog resources
Manage permissions for specific databases or tables, in addition to fine-grained data access.

Databases
Select one or more databases.

Choose databases ▾ Load more

minhnguyetdb X
734081627824

Tables - optional
Select one or more tables.

Choose tables ▾ Load more

All tables X
734081627824

Data filters - optional
Select one or more data filters.

Choose data filters ▾ Load more Create new

Manage data filters

Table permissions

Table permissions
Choose specific access permissions to grant.

<input checked="" type="checkbox"/> Select	<input checked="" type="checkbox"/> Insert	<input checked="" type="checkbox"/> Delete	<input checked="" type="checkbox"/> Super
<input checked="" type="checkbox"/> Describe	<input checked="" type="checkbox"/> Alter	<input checked="" type="checkbox"/> Drop	This permission is the union of all the individual permissions to the left, and supersedes them.

Grantable permissions
Choose the permission that may be granted to others.

<input type="checkbox"/> Select	<input type="checkbox"/> Insert	<input type="checkbox"/> Delete	<input type="checkbox"/> Super
<input type="checkbox"/> Describe	<input type="checkbox"/> Alter	<input type="checkbox"/> Drop	This permission allows the principal to grant any of the permissions to the left, and supersedes those grantable permissions.

Data permissions

<input checked="" type="radio"/> All data access Grant access to all data without any restrictions.	<input type="radio"/> Column-based access Grant data access to specific columns only.
--------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------

[Cancel](#) [Grant](#)

⇒ Kết quả thành công

AWS Lake Formation > Permissions

Too many permissions? Filter by database or table. In the navigation page, choose **Databases** or **Tables**. Then choose a database or table, and on the **Actions** menu, choose **View Permissions**.

Data permissions (8)

Principal	Principal type	Resource type	Database	Table	Resource	Catalog
<input type="radio"/> AWSGlueServiceRole-SalesHocVienCrawler	IAM role	Table	minhnguyetdb	ALL_TABLES	ALL_TABLES	73408162782
<input type="radio"/> AWSGlueServiceRole-SalesHocVienCrawler	IAM role	Column	minhnguyetdb	ALL_TABLES	Included: All	73408162782
<input type="radio"/> IAMAllowedPrincipals	Group	Database	mydatabase	-	mydatabase	73408162782
<input type="radio"/> IAMAllowedPrincipals	Group	Database	minhnguyetdb	-	minhnguyetdb	73408162782
<input type="radio"/> minhnguyet-admin-user	IAM user	Catalog	-	-	-	-
<input type="radio"/> minhnguyet-admin-user	IAM user	Database	minhnguyetdb	-	minhnguyetdb	73408162782

Bước 7: Run Crawler

The screenshot shows the AWS Lambda console with the 'Crawlers' tab selected. At the top, there are buttons for 'Add crawler', 'Run crawler', 'Action', and a search bar labeled 'Filter by tags and attr'. Below this, a table lists a single crawler:

<input checked="" type="checkbox"/>	Name	Schedule
<input checked="" type="checkbox"/>	SalesCrawlerHocVien	

⇒ Kết quả thành công

The screenshot shows the AWS Lambda console with the 'Crawlers' tab selected. At the top, there are buttons for 'Add crawler', 'Run crawler', 'Action', and a search bar labeled 'Filter by tags and attributes'. Below this, a table displays crawler details:

<input type="checkbox"/>	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
<input type="checkbox"/>	SalesCrawlerHocVien		Ready	Logs	57 secs	57 secs	0	1

The screenshot shows the AWS Lake Formation console with the 'Tables' tab selected. At the top, there are buttons for 'Actions' and 'Create table using template'. Below this, a table lists existing tables:

<input checked="" type="checkbox"/>	Name	Database	Governance	Owner account ...	Shared resources
<input checked="" type="checkbox"/>	hocvien	minhnguyetdb	-	734081627824	-

AWS Lake Formation > Tables > hocvien

hocvien Version 0 (Current version) ▾

Table details

Database	Description
minhnguyetdb	-
Location	Data format
s3://minhnguyet-bucket/data/sales/hocvien/	csv
Connection	Last updated
-	Thursday, October 13, 2022 at 5:26 AM UTC

► Advanced table properties

Schema

Find Columns

#	Column Name	▼	Data type	▼	Partition key	Comment	LF-Tags
1	id_hv		bigint		-	-	-
2	họ đệm		string		-	-	-
3	tên		string		-	-	-
4	gioitinh		string		-	-	-
5	sdt		string		-	-	-
6	email		string		-	-	-
7	trinhdo		string		-	-	-

Bước 8: Truy vấn dữ liệu thông qua Athena

Amazon Athena > Query editor > Manage settings

Manage settings

Query result location and encryption

Location of query result - *optional*

Enter an S3 prefix in the current region where the query result will be saved as an object.

 s3://minhnguyet-bucket

Expected bucket owner - *optional*

Specify the AWS account ID that you expect to be the owner of your query results output location bucket.

 Enter AWS account ID

Assign bucket owner full control over query results

Enabling this option grants the owner of the S3 query results bucket full control over the query results. This means that if your query result location is owned by another account, you grant full control over your query results to the other account.

Encrypt query results

The screenshot shows the AWS Data Catalog interface. On the left, under 'Data', the 'Data source' is set to 'AwsDataCatalog' and the 'Database' is 'minhnguyetdb'. In the center, under 'Tables and views', there is a list of tables. One table, 'hocvien', is selected and expanded, showing its columns: id_hv (bigint), họ đệm (string), tên (string), gioitinh (string), sdt (string), email (string), and trinhdo (string). On the right, the 'Query 1' tab is active, displaying the SQL query 'select * from hocvien'. Below the query, the status is 'SQL Ln 1, Col 13' with buttons for 'Run again', 'Explain', 'Cancel', and 'Save'. The 'Completed' status is shown under 'Query results'. At the bottom, it says 'Results (100+)' and 'Search rows'.

Data

Data source

AwsDataCatalog

Database

minhnguyetdb

Tables and views

Create ▾

Filter tables and views

Tables (1)

hocvien

id_hv	bigint	⋮
họ đệm	string	⋮
tên	string	⋮
gioitinh	string	⋮
sdt	string	⋮
email	string	⋮
trinhdo	string	⋮

Query 1

1 | select * from hocvien

SQL Ln 1, Col 13

Run again Explain Cancel Save

Completed

Query results Query stats

Results (100+)

Search rows

⇒ Kết quả truy vấn thu được

Query results		Query stats					
Completed		Time in queue: 110 ms Run time: 1.28 sec Data scanned: 213.06 KB					
Results (100+)							
#	id_hv	họ đệm	tên	gioitinh	sdt	email	trinhdo
1	72019001	Mi Xa	Tran	Female	917520969	mixa.brvt@gmail.com	Kaiwa 1
2	72019002	Hoang Thi	Phuong	Female	842323225	phuongnguyenad1093@gmail.com	Kaiwa 3
3	72019003	Trần Việt	Hạnh	Female	914297679	viethanh.dn@gmail.com	Kaiwa 1
4	72019004	Nguyễn Thái	Hòa	Female	906220311	thaohoantu@gmail.com	Kaiwa 3
5	72019005	Duy	Thành	Male	977701689	Duythanhhdhc46@gmail.com	Kaiwa 5
6	72019006	Duy	Bùi	Male	911263017	Builetuduy@gmail.com	Kaiwa 2
7	72019007	Trần Vĩnh	Hội	Male	386330057	vinhhoi2005@gmail.com	Kaiwa 4
8	72019008	Phạm Khánh	Ngọc	Female	914869998	khanhngoc0208@gmail.com	Kaiwa 5
9	72019009	Trần minh	trung	Male	869690927	tmtrung20101993@gmail.com	Kaiwa 1

Bước 9: Thực hiện truyền dữ liệu trực tiếp thời gian thực của bộ phận Sales sử dụng Kinesis Firehose

- Tạo delivery stream

Amazon Kinesis > Delivery streams > Create delivery stream

Create a delivery stream Info

▶ **Amazon Kinesis Data Firehose: How it works**

Choose source and destination
Specify the source and the destination for your delivery stream. You cannot change the source and destination of your delivery stream once it has been created.

Source Info
Direct PUT

Destination Info
Amazon S3

Destination settings [Info](#)

Specify the destination settings for your delivery stream.

S3 bucket

[Browse](#)
[Create](#)

Format: s3://bucket

Dynamic partitioning [Info](#)

Dynamic partitioning enables you to create targeted data sets by partitioning streaming S3 data based on partitioning keys. You can partition your source data with inline parsing and/or the specified AWS Lambda function. You can enable dynamic partitioning only when you create a new delivery stream. You cannot enable dynamic partitioning for an existing delivery stream. Enabling dynamic partitioning incurs additional costs per GiB of partitioned data. For more information, see [Kinesis Data Firehose pricing](#).

Disabled

Enabled

S3 bucket prefix - optional

By default, Kinesis Data Firehose appends the prefix "YYYY/MM/dd/HH" (in UTC) to the data it delivers to Amazon S3. You can override this default by specifying a custom prefix that includes expressions that are evaluated at runtime.

data/sales/real-time-hocvien/

You can repeat the same keys in your S3 bucket prefix. Maximum S3 bucket prefix characters: 1024.

⇒ Kết quả thành công

Amazon Kinesis > Delivery streams

Delivery streams (1)

C
 Delete

Name	▲	Status	▼	Creation time	▼	Source	▼	Data transfo...	▼	Desti
PUT-S3-ETPR4	Active	October 13, 20...		Direct PUT		Not enabled		Amaz...		

- Cấp quyền dữ liệu

AWS Lake Formation > Grant permissions

Grant data permissions

Principals

IAM users and roles
Users or roles from this AWS account.

SAML users and groups
SAML users and group or QuickSight ARNs.

External accounts
AWS accounts or AWS organizations outside of this account.

IAM users and roles
Add one or more IAM users or roles.

Choose IAM principals to add ▾

KinesisFirehoseServiceRole-PUT-S3-E-ap-southeast-1-1665639761196 X
Role

LF-Tags or catalog resources

Resources matched by LF-Tags (recommended)
Manage permissions indirectly for resources or data matched by a specific set of LF-Tags.

Named data catalog resources
Manage permissions for specific databases or tables, in addition to fine-grained data access.

Databases
Select one or more databases.

Choose databases ▾ Load more

minhnguyetdb X
734081627824

Tables - optional
Select one or more tables.

Choose tables ▾ Load more

All tables X
734081627824

Data filters - optional
Select one or more data filters.

Choose data filters ▾ Load more Create new

Manage data filters

Table permissions

Table permissions
Choose specific access permissions to grant.

<input type="checkbox"/> Select	<input type="checkbox"/> Insert	<input type="checkbox"/> Delete	<input checked="" type="checkbox"/> Super
<input type="checkbox"/> Describe	<input type="checkbox"/> Alter	<input type="checkbox"/> Drop	This permission is the union of all the individual permissions to the left, and supersedes them.

- Test with demo data

Amazon Kinesis > Delivery streams > PUT-S3-ETPR4

PUT-S3-ETPR4 Info

Delivery stream details

Status
Active

Source
Direct PUT

▼ Test with demo data Info
Ingest simulated data to test the configuration of your delivery stream. Standard Amazon Kinesis Data Firehose delivery streams are charged at a rate of \$1,000 per GB of data ingested.

This test runs a script in your browser to put demo data in your Kinesis Data Firehose delivery stream. The demo data is generated from the following JSON object:

```
1 {
2   "TICKER_SYMBOL": "QXZ",
3   "SECTOR": "HEALTHCARE",
4   "CHANGE": -0.05,
5   "PRICE": 84.51
6 }
```

Step 1

Start sending demo data to your delivery stream. If you already have data streaming to this destination, demo data will be added to it.

✓ **Sending demo data**

Step 2

Stop sending demo data to your delivery stream after you've concluded your test to stop incurring usage charges.

Stop sending demo data

⇒ Kết quả nhận được

Amazon S3 > Buckets > minhnguyet-bucket > data/ > sales/ > real-time-hocvien/ > 2022/ > 10/ > 13/ > 06/

06/

Objects (8)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Actions

- C
- Copy S3 URI
- Copy URL
- Download
- Open
- Delete
- Actions
- Create folder
- Upload

Find objects by prefix

Name	Type	Last modified	Size	Storage class
PUT-S3-ETPR4-2-2022-10-13-06-02-56-07ead7bb-5a57-4ed1-8172-87a0b63e82de	-	October 13, 2022, 13:07:57 (UTC+07:00)	4.0 KB	Standard
PUT-S3-ETPR4-2-2022-10-13-06-07-59-a637517-ec6d-45cb-a042-74af115bc0f6	-	October 13, 2022, 13:13:00 (UTC+07:00)	3.6 KB	Standard
PUT-S3-ETPR4-2-2022-10-13-06-13-05-81212a7d-7a10-4d42-97be-8ebf524c20de	-	October 13, 2022, 13:18:07 (UTC+07:00)	4.0 KB	Standard

- Add crawler (tương tự như bước 5)

Crawler info

Name: SalesRealTimeHocVien
Tags: -

Data stores

Data store: S3
Include path: s3://minhnguyet-bucket/data/sales/real-time-hocvien
Connection
Exclude patterns

IAM role

IAM role: arn:aws:iam::734081627824:role/service-role/AWSGlueServiceRole-SalesRealTimeHocVienCrawler

Schedule

Schedule: Run on demand

Output

Database: minhnguyetdb
Prefix added to tables (optional)
Table threshold (optional)
Create a single schema for each S3 path: false
Table level (optional)
Configuration options

Back Finish

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawler "SalesRealTimeHocVien" completed and made the following changes: 1 tables created, 0 tables updated. See the tables created in database [minhnguyetdb](#).

User
Showing: 1 - 2 <

<input type="checkbox"/>	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables created
<input type="checkbox"/>	SalesCrawlerHocVien		Ready	Logs	57 secs	57 secs	0	1
<input type="checkbox"/>	SalesRealTimeHocVien		Ready	Logs	51 secs	51 secs	0	1

Data permissions (13)

[Revoke](#)
[Grant](#)

Filter permissions by property or value
< 1 ... >

Principal	▲	Principal type ▼	Resource type ▼	Database ▼	Table ▼	Resource ▼
<input type="checkbox"/> AWSGlueServiceRole-SalesHocVienCrawler		IAM role	Table	minhnguyetdb	hocvien	hocvien
<input type="checkbox"/> AWSGlueServiceRole-SalesHocVienCrawler		IAM role	Column	minhnguyetdb	hocvien	Includes
<input type="checkbox"/> AWSGlueServiceRole-SalesHocVienCrawler		IAM role	Table	minhnguyetdb	ALL_TABLES	ALL_TABLES
<input type="checkbox"/> AWSGlueServiceRole-SalesHocVienCrawler		IAM role	Column	minhnguyetdb	ALL_TABLES	Includes
<input type="checkbox"/> AWSGlueServiceRole-SalesRealTimeHocVienCrawler		IAM role	Table	minhnguyetdb	ALL_TABLES	ALL_TABLES

- Truy vấn real_time_hocvien

The screenshot shows the AWS Data Catalog interface. On the left, the 'Data' sidebar displays 'Data source' set to 'AwsDataCatalog' and 'Database' set to 'minhnguyetdb'. Below these are sections for 'Tables and views' (with a 'Create' button) and a search bar for filtering tables and views. A dropdown menu shows 'Tables (2)': 'hocvien' and 'real_time_hocvien'. The 'real_time_hocvien' table is expanded, showing columns: 'change' (double), 'price' (double), 'ticker_symbol' (string), 'sector' (string), 'year' (string (Partitioned)), 'month' (string (Partitioned)), 'day' (string (Partitioned)), and 'hour' (string (Partitioned)).

On the right, the main area is titled 'Query 1' with the SQL query: 'select * from real_time_hocvien'. Below the query, the status is 'SQL Ln 1, Col 32' with buttons for 'Run again' (orange), 'Explain', and 'Cancel'. The 'Query results' tab is selected, showing a green bar with 'Completed'. The 'Results (11)' section shows a table header with columns: '#', 'change', 'price', and 'ticl' (partially visible). A search bar for rows is also present.

⇒ Kết quả nhận được:

Results (11)											<input type="button" value="Copy"/>	<input type="button" value="Download results"/>
#	change	price	ticker_symbol	sector	year	month	day	hour				
1	5.26	203.2	QAZ	FINANCIAL	2022	10	13	06				
2	-5.13	99.39	DFT	RETAIL	2022	10	13	06				
3	14.54	788.6	AMZN	TECHNOLOGY	2022	10	13	06				
4	-8.1	90.5	NFS	ENERGY	2022	10	13	07				
5	-1.46	97.54	NFLX	TECHNOLOGY	2022	10	13	05				
6	-1.04	113.44	QXZ	HEALTHCARE	2022	10	13	06				
7	-1.64	92.75	NFLX	TECHNOLOGY	2022	10	13	06				
8	0.32	18.49	MMB	ENERGY	2022	10	13	06				
9	1.96	118.2	IOP	TECHNOLOGY	2022	10	13	06				
10	0.17	99.61	DFT	RETAIL	2022	10	13	06				
11	-2.32	66.07	TGT	RETAIL	2022	10	13	05				

Bước 10: Lake Formation and Glue ETL

- Tạo database

Thực hiện tương tự bước 4.

- Add job

Configure the job properties

Name

IAM role ⓘ

Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job. [Create IAM role](#).

Type

Glue version

This job runs

A proposed script generated by AWS Glue ⓘ

An existing script that you provide

A new script to be authored by you

Script file name

S3 path where the script is stored

Temporary directory ⓘ

Choose a data source

Filter by attributes or search by keyword

Name	Database	Location	Classification
<input checked="" type="radio"/> real_time_hocvien	minhnguyetdb	s3://minhnguyet-bucket/data/sal...	json
<input type="radio"/> hocvien	minhnguyetdb	s3://minhnguyet-bucket/data/sal...	csv

Choose a data target

Create tables in your data target
 Use tables in the data catalog and update your data target

Data store

Amazon S3

Format

Parquet

Connection

- Select one -

Add connection

Target path

s3://minhnguyet-bucket/verified/RealTimeParquest-ver

[Back](#) [Next](#)

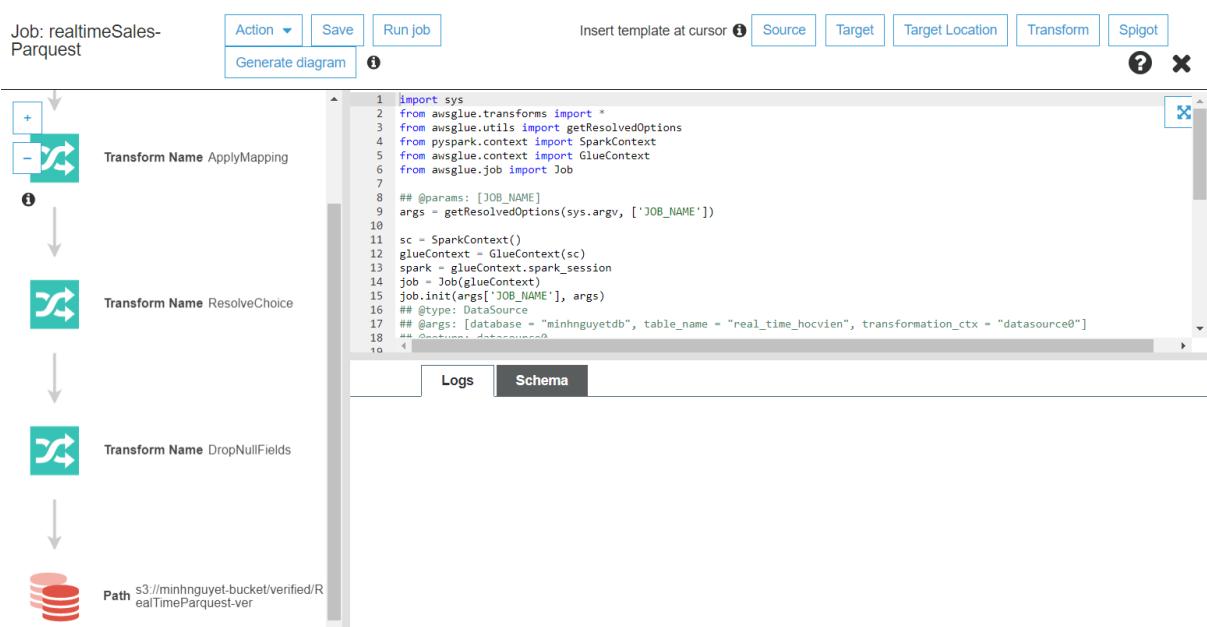
Output Schema Definition

Verify the mappings created by AWS Glue. Change mappings by choosing other columns with **Map to target**. You can **Clear** all mappings and **Reset** to default AWS Glue mappings. AWS Glue generates your script with the defined mappings.

Source	Target			
Column name	Data type	Map to target	Column name	Data type
change	double	change	change	double
price	double	price	price	double
ticker_symbc	string	ticker_symbol	ticker_symbol	string
sector	string	sector	sector	string
year	string	year	sector	string
month	string	month	sector	string
day	string	day	month	string
hour	string	hour	day	string
			hour	string

[Add column](#) [Clear](#) [Reset](#)

⇒ Kết quả thu được



- Run job

Kết quả của Tab Logs khu Run job

Amazon S3 > Buckets > minhnguyet-bucket > verified/

verified/

Objects Properties

Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions.

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified
<input type="checkbox"/>	RealTimeParquest-ver_\$folder\$	-	October 13, 2022, 16:20:44 (UTC+07:00)
<input type="checkbox"/>	RealTimeParquest-ver/	Folder	-

Amazon S3 > Buckets > minhnguyet-bucket > verified/ > RealTimeParquest-ver/

RealTimeParquest-ver/

Objects Properties

Objects (11)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	
<input type="checkbox"/>	part-00000-7822f7d4-eeac-4fd6-a00b-6bee2eaca70c-c000.snappy.parquet	parquet	October 13, 2022, 16:20:36 (UTC+07:00)	2.3 KB	<input type="checkbox"/>
<input type="checkbox"/>	part-00001-7822f7d4-eeac-4fd6-a00b-6bee2eaca70c-c000.snappy.parquet	parquet	October 13, 2022, 16:20:36 (UTC+07:00)	2.7 KB	<input type="checkbox"/>
<input type="checkbox"/>	part-00002-7822f7d4-eeac-4fd6-a00b-6bee2eaca70c-c000.snappy.parquet	parquet	October 13, 2022, 16:20:36 (UTC+07:00)	2.9 KB	<input type="checkbox"/>
<input type="checkbox"/>	part-00003-7822f7d4-eeac-4fd6-a00b-6bee2eaca70c-c000.snappy.parquet	parquet	October 13, 2022, 16:20:36 (UTC+07:00)	2.9 KB	<input type="checkbox"/>
<input type="checkbox"/>	part-00004-7822f7d4-eeac-4fd6-a00b-6bee2eaca70c-c000.snappy.parquet	parquet	October 13, 2022, 16:20:36 (UTC+07:00)	2.8 KB	<input type="checkbox"/>
<input type="checkbox"/>	part-00005-7822f7d4-eeac-4fd6-a00b-6bee2eaca70c-c000.snappy.parquet	parquet	October 13, 2022, 16:20:36 (UTC+07:00)	2.9 KB	<input type="checkbox"/>
<input type="checkbox"/>	part-00006-7822f7d4-eeac-4fd6-a00b-6bee2eaca70c-c000.snappy.parquet	parquet	October 13, 2022, 16:20:36 (UTC+07:00)	2.9 KB	<input type="checkbox"/>
<input type="checkbox"/>	part-00007-7822f7d4-eeac-4fd6-a00b-6bee2eaca70c-c000.snappy.parquet	parquet	October 13, 2022, 16:20:36 (UTC+07:00)	2.0 KB	<input type="checkbox"/>

Editor Recent queries Saved queries Settings

Data C <

Query 1 x Query 2 x

1 select * from realtimeparquest_ver

Data source: AwsDataCatalog

Database: verified

Tables and views Create ▾ ⚙

Filter tables and views

▼ Tables (1) < 1 >

realtimeparquest_ver

change	double	⋮
price	double	⋮
ticker_symbol	string	⋮
sector	string	⋮
year	string	⋮

SQL Ln 1, Col 35

Run again Explain ⚡ Cancel

Query results Query stats

Completed

This screenshot shows the AWS Glue Data Catalog Editor interface. At the top, there are tabs for 'Editor' (which is selected), 'Recent queries', 'Saved queries', and 'Settings'. Below the tabs, the 'Data' section is visible, showing a 'Data source' dropdown set to 'AwsDataCatalog' and a 'Database' dropdown set to 'verified'. In the 'Tables and views' section, there's a 'Create' button and a search bar. A 'Tables (1)' section is expanded, showing a single table named 'realtimeparquest_ver' with five columns: 'change' (double), 'price' (double), 'ticker_symbol' (string), 'sector' (string), and 'year' (string). To the right of the table list, there's a sidebar for the currently selected query. The sidebar shows the query 'select * from realtimeparquest_ver' with two runs: 'Query 1' (green checkmark) and 'Query 2' (orange checkmark). The status for both runs is 'Completed'. Below the sidebar, there are buttons for 'Run again', 'Explain ⚡', and 'Cancel'. There are also tabs for 'Query results' and 'Query stats'.

⇒ Kết quả truy vấn

Query results		Query stats	
🕒 Completed		Time in queue: 120 ms Run time: 1.05 sec Data scanned: 14.29 KB	
Results (100+)			
<input type="text"/> Search rows			Copy Download results
# ▾ change ▾ price ▾ ticker_symbol ▾ sector ▾ year ▾ month ▾ day ▾ hour ▾			
1	0.32	18.49	MMB ENERGY 2022 10 13 06
2	-0.21	5.88	KIN ENERGY 2022 10 13 06
3	-8.48	126.19	QXZ HEALTHCARE 2022 10 13 06
4	10.44	187.02	TBV HEALTHCARE 2022 10 13 06
5	-3.76	41.71	SAC ENERGY 2022 10 13 06
6	1.24	54.65	RFV FINANCIAL 2022 10 13 06
7	0.26	5.94	DEG ENERGY 2022 10 13 06
8	2.01	43.72	SAC ENERGY 2022 10 13 06
9	-1.01	21.75	ABC RETAIL 2022 10 13 06
10	-0.01	15.47	JKL TECHNOLOGY 2022 10 13 06

CHƯƠNG 7. XÂY DỰNG CÁC BÁO CÁO PHÂN TÍCH THỐNG KÊ

7.1. Xác định BI User

Những người sẽ sử dụng hệ thống thông tin báo cáo tình hình kinh doanh (BI users) trong doanh nghiệp được mô tả trong bảng dưới đây.

Chức vụ	Nghệ vụ	Quản lý	Cấp ra quyết định
Trưởng phòng Hành chính - Nhân sự	Phát triển và quản lý các kế hoạch và quy trình nguồn nhân lực của công ty. Lập kế hoạch, tổ chức và kiểm soát các hoạt động của bộ phận Nhân sự. Thúc đẩy vào việc phát triển các mục tiêu và hệ thống của bộ phận Nhân sự.	Quản lý nhân sự và chính sách chăm sóc nhân viên	Cấp chiến lược
Trưởng phòng marketing	Hoạch định chiến lược, kế hoạch, giải pháp và tổ chức thực hiện hoạt động Marketing của công ty. Tổ chức thực hiện các hoạt động và chương trình nghiên cứu thị trường.	Quản lý các kế hoạch marketing	Cấp chiến lược

Trưởng phòng kinh doanh	<p>Xác định định hướng kinh doanh hướng tới sự phát triển và lợi nhuận của doanh nghiệp.</p> <p>Chịu trách nhiệm về hiệu quả bán hàng của sản phẩm.</p> <p>Phát triển chiến lược nhằm tạo ra các cơ hội kinh doanh phù hợp với mục đích tăng trưởng doanh thu cho doanh nghiệp.</p> <p>Xây dựng ngân sách cho các kế hoạch ngắn hạn và dài hạn liên quan đến lợi nhuận và chi tiêu của doanh nghiệp.</p>	Quản lý nhân sự đội ngũ kinh doanh để hoàn thành tốt KPI bán hàng	Cấp chiến lược
Trưởng phòng kế toán	Kiểm tra chứng từ...Hạch toán thu nhập, chi phí, lợi nhuận...	Quản lý tài chính kế toán, đầu tư thông kê và xử lý chứng từ liên quan.	Cấp chiến thuật

Bảng 7.1 BI User của công ty

7.2. Xây dựng các báo cáo phân tích thống kê (dashboard).

7.2.1. Bảng câu hỏi

Trong quy trình kinh doanh cũng như quy trình quản lý học tập, người dùng BI thường đặt ra những câu hỏi phục vụ quá trình ra quyết định nhằm đạt hiệu quả cao nhất: Lượng khóa học bán ra nhiều hay ít, có ổn định không? Giáo viên đã đủ đáp ứng nhu cầu lớp học hay chưa? Tình hình biến động của doanh thu như thế nào? ... Cụ thể

Trưởng phòng kinh doanh	<ul style="list-style-type: none"> - Tổng doanh thu của công ty theo khóa học, theo nhân viên Sale, qua từng tháng/quý/năm? - Doanh thu và tỉ lệ doanh thu theo loại khóa học (khóa combo, khóa học lẻ...), theo hình thức lớp học (1-1, 1-4) ?
Trưởng phòng marketing	<ul style="list-style-type: none"> - Ảnh hưởng của các chương trình khuyến mãi? - Lượng học viên đăng ký khóa học vào các đợt khuyến mãi của công ty?
Trưởng phòng hành chính-nhân sự	<ul style="list-style-type: none"> - Số học viên theo học và lượng lớp giảng dạy của giáo viên là bao nhiêu? - Tỉ lệ giáo viên theo giới tính, quốc tịch? - Trạng thái các lớp qua thời gian? - Số lượng lớp học và giáo viên theo ngày học (T2-T4-T6, T3-T5-T7) ? - Danh sách nhân viên sale có hiệu suất cao nhất (số lượng khóa học đã bán, doanh thu mang lại) ? - Doanh thu từng tháng của từng nhân viên sale?
Trưởng phòng kế toán	<ul style="list-style-type: none"> - Thống kê số tiền nhận qua từng tài khoản ngân hàng

Bảng 7.2 Bảng câu hỏi

7.2.2. Lập báo cáo

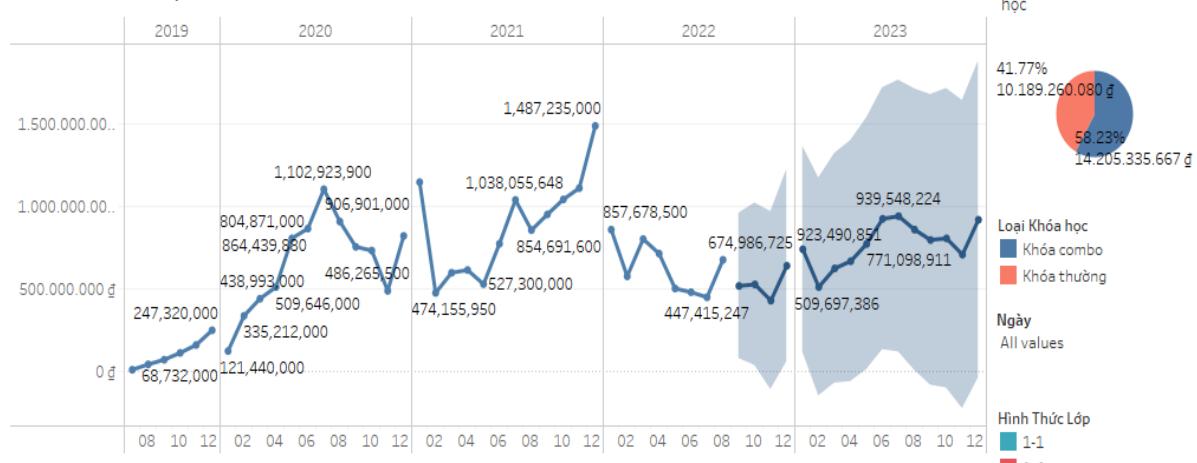
Sử dụng công cụ hỗ trợ phân tích và trực quan hóa dữ liệu Tableau, nhóm đã xây dựng các báo cáo dưới đây.

7.2.2.1. Báo cáo tổng quan tình hình kinh doanh

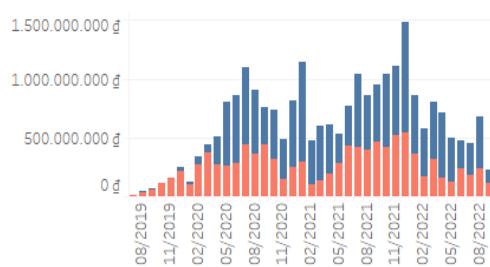
Tổng quan

Số lượng Khóa học	Số lượng NV Sale	Số lượng đăng ký	Tổng số Giáo viên	Tổng số Học viên	Tổng số lớp học	Tổng Doanh thu
65	34	2,800	245	2,800	2,800	24.394.595.747 ₫

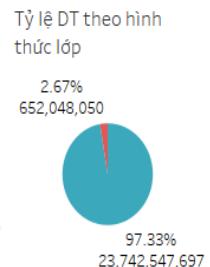
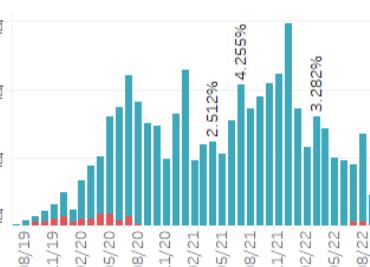
Doanh thu & dự báo doanh thu



Doanh thu theo từng loại khóa học



Doanh thu theo hình thức lớp học

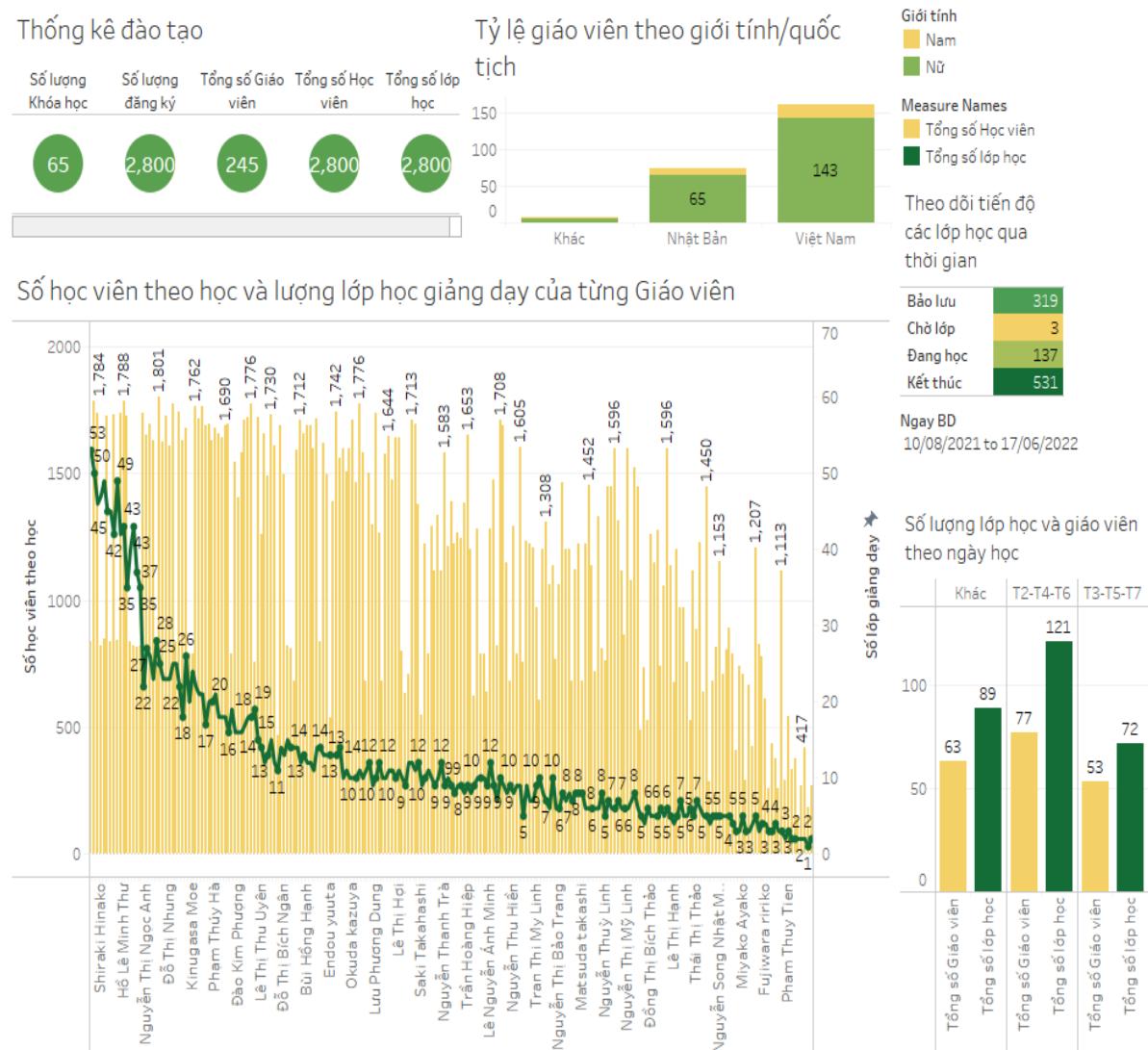


Hình 7.1 Báo cáo tổng quan tình hình kinh doanh

Nhìn vào báo cáo tổng quan tình hình kinh doanh, người dùng có thể biết được các thông số về nhân viên, giáo viên, lớp học tại công ty cũng như tổng doanh thu tính tới thời điểm hiện tại. Doanh thu được thống kê theo từng tháng, có sự so sánh về tỷ lệ doanh thu trên từng loại hình dịch vụ, từ đó đưa ra giải pháp hoặc kế hoạch kinh doanh phù hợp. Ngoài ra, từ dữ liệu doanh thu trước đó, báo cáo đưa ra con số dự đoán doanh thu 16 tháng tiếp theo (từ tháng 9/2022 cho đến tháng 12/2023).

Bên cạnh đó, người dùng có thể linh hoạt chọn hiển thị các con số thống kê theo từng thời điểm hoặc theo từng khoảng thời gian khác nhau.

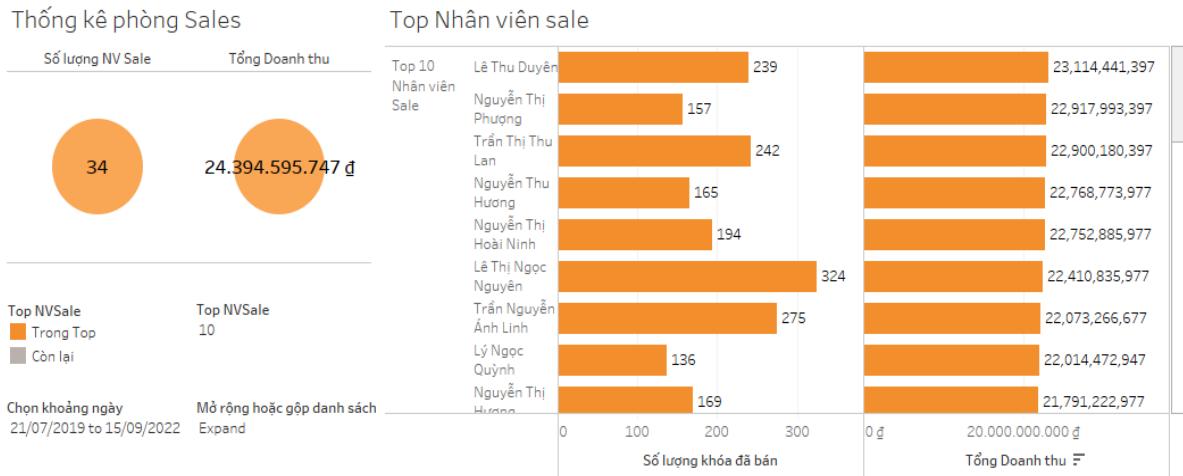
7.2.2.2. Báo cáo quản lý đào tạo



Hình 7.2 Báo cáo quản lý đào tạo

Báo cáo quản lý đào tạo cung cấp góc nhìn tổng quan về hoạt động đào tạo của công ty. Người dùng có thể theo dõi tiến độ của các lớp học tại từng thời điểm nhất định. Số học viên theo học và số lớp học cho biết các con số chi tiết về hoạt động giảng dạy phản ánh kết quả làm việc của cả đội ngũ giáo viên cũng như của từng người cụ thể, từ đó đưa ra chính sách khen thưởng/kỷ luật phù hợp. Kết quả so sánh từ biểu đồ số lượng lớp học và số giáo viên theo ngày học hỗ trợ người dùng đưa ra các quyết định về nhân sự, dịch vụ cần thiết để đáp ứng nhu cầu của người học.

7.2.2.3. Báo cáo doanh thu phòng Sales



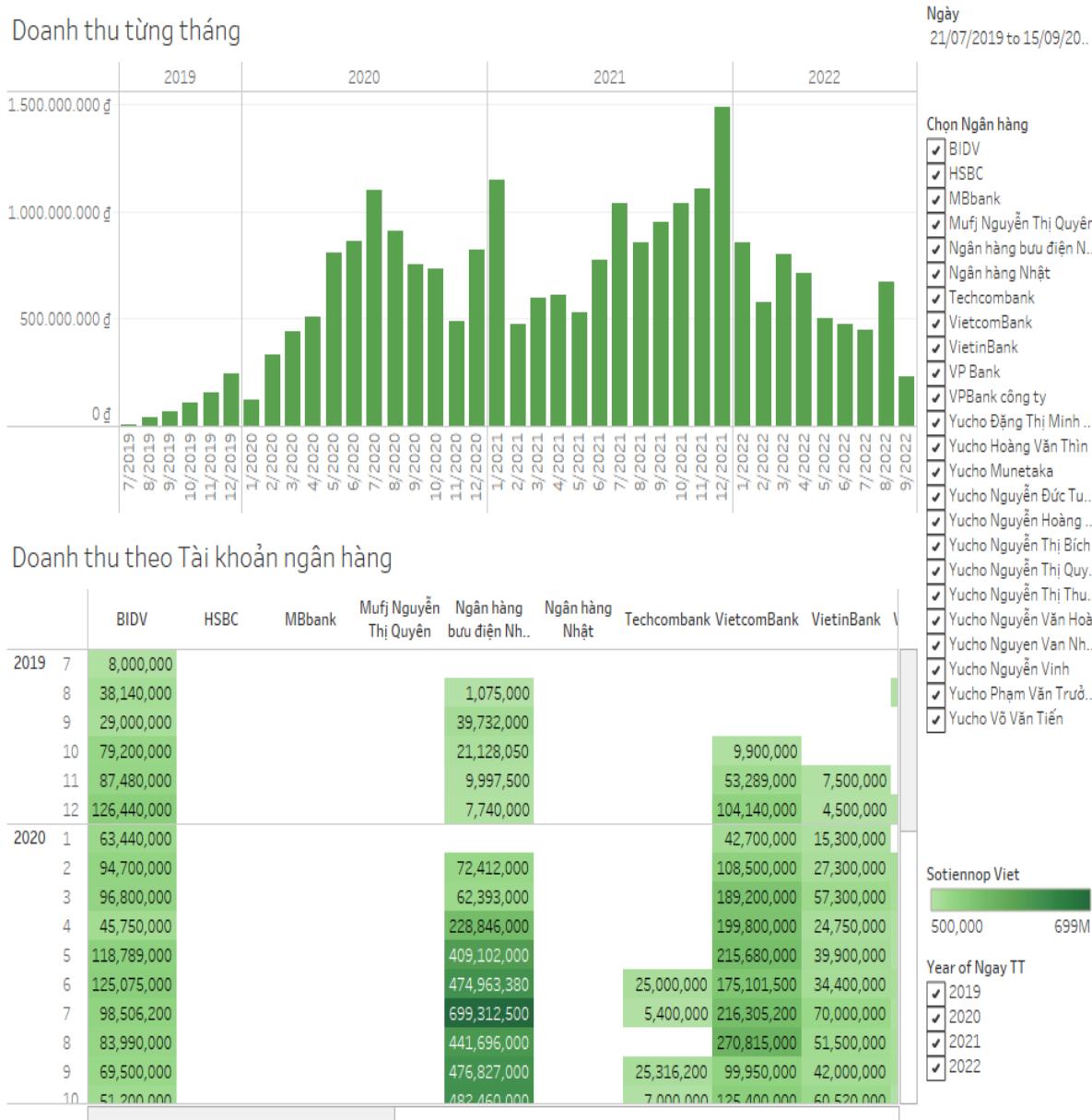
Chi tiết doanh thu từng tháng của Nhân viên Sale

	01/2021	02/2021	03/2021	04/2021	05/2021	06/2021	07/2021	08/2021	09/2021
Đặng Kim Oanh	223,957,945	100,445,100	112,300,000	63,200,000	113,700,000	203,575,000	157,331,800	212,151,000	199,981,500
Đặng Minh Trí	52,065,000	16,600,000	19,700,000	20,800,000	61,200,000	139,000,000	53,310,000	78,425,000	86,765,000
Đặng Thị Lam Trang	904,944,045	415,942,900	484,594,000	539,880,000	448,500,000	654,535,000	962,200,648	707,516,600	845,526,500
Đặng Thị Thanh Hoài	922,514,045	432,153,650	493,224,000	553,280,000	467,000,000	714,415,000	991,975,648	737,876,600	893,446,500
Đặng Thị Thanh Kim	978,524,045	454,153,650	520,194,000	572,280,000	474,950,000	711,635,000	953,225,648	750,916,600	864,801,500
Hà Ánh Minh	724,664,045	307,362,900	400,524,000	387,480,000	409,850,000	624,040,000	833,475,600	656,706,000	812,906,500
Lê Minh Huyền	550,851,100	270,990,750	327,794,000	390,680,000	270,050,000	416,439,000	639,368,848	377,560,600	526,720,000
Lê Thị Ngọc Nguyên	1,081,934,045	472,155,950	571,594,000	602,780,000	514,800,000	744,285,000	1,003,980,648	761,416,600	919,686,500
Lê Thu Duyên	1,065,839,045	454,153,650	557,824,000	602,780,000	514,800,000	772,515,000	1,025,960,648	854,691,600	950,186,500
Lý Ngọc Quỳnh	981,439,045	434,153,650	501,194,000	546,780,000	461,000,000	652,535,000	970,200,648	741,516,600	860,646,500
Nguyễn Ngọc Trâm	299,982,945	127,561,900	174,900,000	174,000,000	177,300,000	285,660,000	356,490,600	383,511,000	361,981,500
Nguyễn Thanh Khuê	52,065,000	16,600,000	19,700,000	20,800,000	61,200,000	139,000,000	53,310,000	78,425,000	86,765,000
Nguyễn Thị Ánh Đào	967,219,045	435,645,200	518,794,000	553,380,000	393,750,000	654,635,000	890,545,648	678,856,600	780,766,500
Nguyễn Thị Hoài Ninh	1,062,434,045	472,155,950	576,224,000	612,780,000	491,500,000	761,515,000	1,015,460,648	806,791,600	937,236,500
Nguyễn Thị Hương	1,071,034,045	455,945,200	564,594,000	595,880,000	485,500,000	736,285,000	995,980,648	761,416,600	906,566,500
Nguyễn Thị Mỹ Linh	418,786,100	162,410,750	246,594,000	258,280,000	248,450,000	366,714,000	538,028,800	340,890,000	516,185,000
Nguyễn Thị Phương	1,050,339,045	454,153,650	557,824,000	602,780,000	498,000,000	772,515,000	1,003,160,648	814,276,600	950,186,500
Nguyễn Thị Huường	934,129,045	452,153,650	504,594,000	592,780,000	474,450,000	724,285,000	953,225,648	714,516,600	858,151,500
Nguyễn Thu Hướng	1,103,034,045	472,155,950	586,724,000	612,780,000	498,000,000	772,515,000	1,003,160,648	780,276,600	948,186,500
Phan Hoàng Long	58,495,000	4,000,000	46,900,000	55,200,000	66,800,000	120,564,000	119,818,800	135,225,000	117,250,000
Phan Thị Khanh Huyền	861,724,045	398,341,900	479,794,000	564,580,000	432,400,000	653,285,000	928,385,648	698,846,600	859,006,500
Phòng Đào tạo	996,004,045	455,945,200	551,224,000	581,880,000	479,000,000	753,515,000	995,160,648	766,376,600	924,116,500
Trần Công Định	359,837,945	140,955,850	179,000,000	163,300,000	255,450,000	379,139,000	378,490,600	404,536,000	438,901,500
Trần Khánh Linh	989,714,045	434,554,950	550,924,000	581,480,000	451,400,000	700,515,000	958,160,648	731,606,600	911,576,500
Trần Mai Vy	255,957,945	116,655,850	119,300,000	80,100,000	126,200,000	211,575,000	165,331,800	212,151,000	213,101,500
Trần Nguyễn Ánh Linh	990,514,045	452,153,650	530,224,000	598,780,000	498,000,000	772,515,000	1,003,160,648	780,276,600	948,186,500
Trần Thị Dung	52,065,000	16,600,000	19,700,000	20,800,000	61,200,000	139,000,000	53,310,000	78,425,000	86,765,000
Tổng	512,217,045	255,441,000	260,000,000	255,400,000	260,000,000	261,205,000	260,000,000	261,211,000	261,211,000

Hình 7.3 Báo cáo doanh thu phòng Sales

Linh hoạt chọn lựa thời gian, người dùng có thể xem danh sách chi tiết doanh thu của nhân viên. Tùy vào các điều chỉnh trên bộ lọc, người dùng có thể nắm bắt được top những nhân viên có thành tích xuất sắc trong hoạt động từ đó áp dụng chính sách lương thưởng, khen chê tương ứng.

7.2.2.4. Báo cáo doanh thu tại phòng kế toán



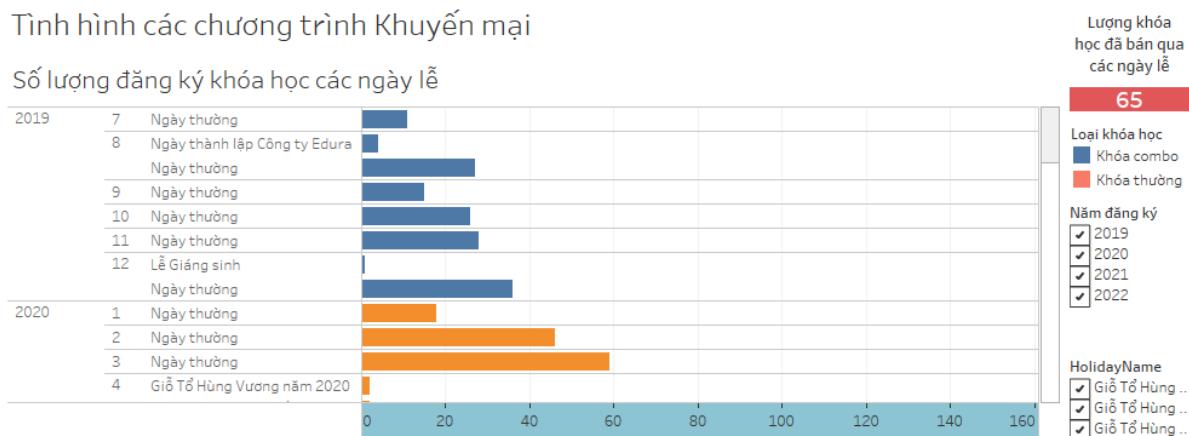
Hình 7.4 Báo cáo doanh thu tại phòng kế toán

Số liệu thể hiện trong báo cáo doanh thu tại phòng kế toán đều có thể được thay đổi linh hoạt theo các điều kiện về: ngày, tháng, năm, loại tài khoản ngân hàng nhận các khoản thanh toán học phí của học viên, từ đó người dùng có thể hạch toán doanh thu, lợi nhuận, quyết toán thuế và có điều chỉnh sổ sách cần thiết.

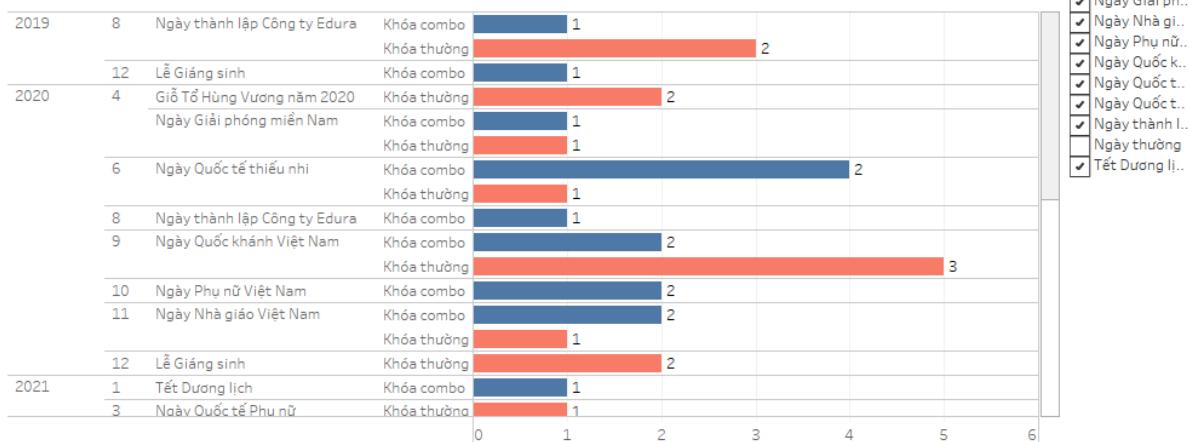
7.2.2.5. Báo cáo lượng khóa học đăng ký vào các đợt khuyến mại

Tình hình các chương trình Khuyến mại

Số lượng đăng ký khóa học các ngày lễ



Khóa học bán chạy nhất trong các ngày lễ



Hình 7.5 Báo cáo lượng khóa học đăng ký vào các đợt khuyến mại

Từ kết quả theo dõi được trên báo cáo lượng khóa học đăng ký vào các đợt khuyến mại, người dùng BI có thể đưa ra các phương án, chiến lược, xây dựng thực hiện các chương trình ưu đãi, khuyến mại cho từng thời điểm hiệu quả hơn.

CHƯƠNG 8. ỨNG DỤNG MACHINE LEARNING TRONG XÂY DỰNG MÔ HÌNH DỰ ĐOÁN HỌC VIÊN TÁI TỤC

8.1. Tổng quan về Machine Learning (Học máy)

8.1.1. Khái niệm Machine Learning

Machine Learning (Học máy) là một nhánh của trí tuệ nhân tạo (AI) và khoa học máy tính, tập trung vào việc sử dụng dữ liệu và thuật toán để học tập cách con người học, dần dần cải thiện độ chính xác của nó. (Theo IBM)

8.1.2. Phân loại

Các nhóm giải thuật học máy:

- Học có giám sát: Máy tính được xem một số mẫu gồm đầu vào (input) và đầu ra (output) tương ứng trước. Sau khi học xong các mẫu này, máy tính quan sát một đầu vào mới và cho ra kết quả.
- Học không giám sát: Máy tính chỉ được xem các mẫu không có đầu ra, sau đó máy tính phải tự tìm cách phân loại các mẫu này và các mẫu mới.
- Học nửa giám sát: Một dạng lai giữa hai nhóm giải thuật trên.
- Học tăng cường: Máy tính đưa ra quyết định hành động (action) và nhận kết quả phản hồi (response/reward) từ môi trường (environment). Sau đó máy tính tìm cách chỉnh sửa cách ra quyết định hành động của mình.

8.1.3. Quy trình xây dựng mô hình machine learning

Quy trình xây dựng mô hình machine learning gồm 5 bước:

- Thu thập dữ liệu: Thu thập dữ liệu để mô hình học
- Chuẩn bị dữ liệu: Xử lý và đưa dữ liệu về định dạng tối ưu, trích chọn đặc trưng hoặc giảm chiều dữ liệu
- Huấn luyện: Tại pha này, thuật toán machine learning thực hiện việc học thông qua các ví dụ đã được thu thập và chuẩn bị từ hai bước trên
- Đánh giá: Kiểm thử mô hình để đánh giá chất lượng của mô hình tốt đến đâu

- Tinh chỉnh: Tinh chỉnh mô hình để tối ưu hiệu quả

8.1.4. Các ứng dụng của học máy

Học máy có ứng dụng rộng khắp trong các ngành khoa học/sản xuất, y tế, giáo dục,... đặc biệt những ngành cần phân tích khôi lượng dữ liệu khổng lồ.

Một số ứng dụng thường thấy:

- Xử lý ngôn ngữ tự nhiên (Natural Language Processing): xử lý văn bản, giao tiếp người – máy, ...
- Nhận dạng (Pattern Recognition): nhận dạng tiếng nói, chữ viết tay, vân tay, thị giác máy (Computer Vision) ...
- Tìm kiếm (Search Engine).
- Chẩn đoán trong y tế: phân tích ảnh X-quang, các hệ chuyên gia chẩn đoán bệnh tự động.
- Tin sinh học: phân loại chuỗi gene, quá trình hình thành gene/protein.
- Vật lý: phân tích ảnh thiên văn, tác động giữa các hạt ...
- Phát hiện gian lận tài chính (financial fraud): gian lận thẻ tín dụng.
- Phân tích thị trường chứng khoán (stock market analysis) Chơi trò chơi: tự động chơi cờ, hành động của các nhân vật ảo.

8.2. Giới thiệu về mô hình cây quyết định (Decision Tree) và thuật toán rừng ngẫu nhiên (Random forest)

Cây quyết định (Decision Tree) và rừng ngẫu nhiên (Random forest) là kỹ thuật được sử dụng rộng rãi trong khai phá dữ liệu, có thể giải quyết được các bài toán phân lớp và hồi quy.

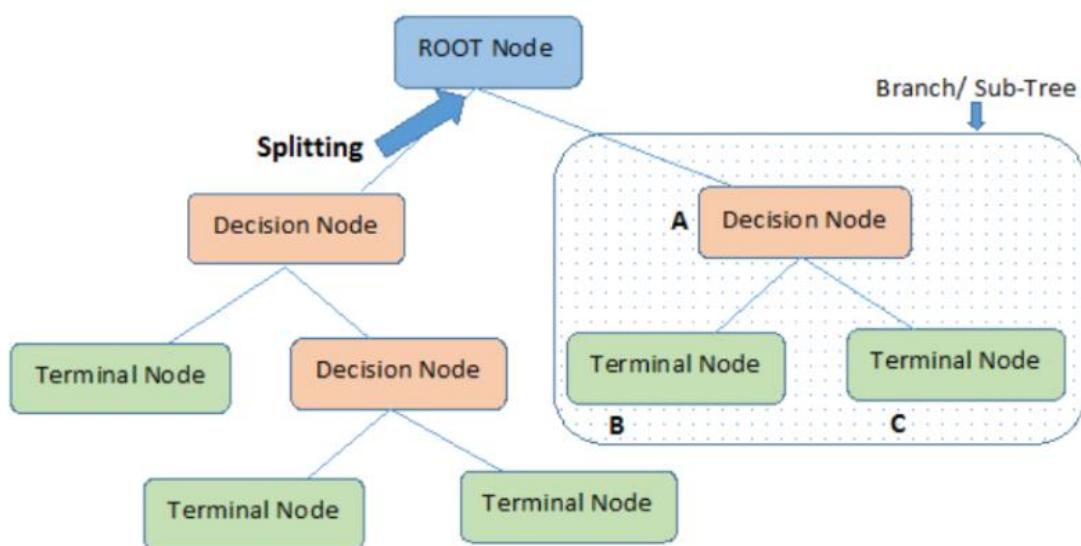
8.2.1. Mô hình Decision Tree (Cây quyết định)

Decision Tree (Cây quyết định) là thuật toán học có giám sát, có thể giải quyết cả bài toán hồi quy và phân lớp.

Trong lĩnh vực máy học, cây quyết định là một kiểu mô hình dự báo (predictive model), nghĩa là một ánh xạ từ các quan sát về một sự vật/hiện tượng tới các kết luận về

giá trị mục tiêu của sự vật/hiện tượng. Mỗi một nút trong (internal node) tương ứng với một biến; đường nối giữa nó với nút con của nó thể hiện một giá trị cụ thể cho biến đó. Mỗi nút lá đại diện cho giá trị dự đoán của biến mục tiêu, cho trước các giá trị của các biến được biểu diễn bởi đường đi từ nút gốc tới nút lá đó.

Decision Tree sử dụng một sơ đồ giống như cấu trúc cây để hiển thị các dự đoán là kết quả của một loạt các phân tách dựa trên tính năng. Nó bắt đầu với một Root Nodes và kết thúc bằng một quyết định của các lá.



Hình 8.1 Cấu trúc của 1 cây quyết định (Nguồn: Decision Tree Algorithms-Machine Learning, Rupika Nimbalkar)

- Root Nodes – Nó là nút hiện diện ở đầu Decision Tree từ nút này, quần thể bắt đầu phân chia theo các đặc điểm khác nhau.
- Các Decision Nodes – các nút chúng ta nhận được sau khi tách các Root Nodes được gọi là Decision Nodes
- Leaf Nodes – các nút không thể tách thêm được gọi là Leaf Nodes hoặc nút đầu cuối
- Sub-tree – giống như một phần nhỏ của đồ thị được gọi là đồ thị con, tương tự như vậy một phần con của Decision Tree này được gọi là Sub-tree.

- Pruning – không có gì khác ngoài việc cắt giảm một số nút để ngừng trang bị quá mức.

Ưu điểm: Một số ưu điểm của cây quyết định là: Đơn giản để hiểu và để giải thích. Yêu cầu chuẩn bị ít dữ liệu. Các kỹ thuật khác thường yêu cầu chuẩn hóa dữ liệu, các biến giả cần được tạo và loại bỏ các giá trị trống. Chi phí sử dụng cây (tức là dữ liệu dự đoán) là logarit trong số điểm dữ liệu được sử dụng để đào tạo cây. Có thể xử lý cả dữ liệu số và dữ liệu phân loại.

Nhược điểm: Người học cây quyết định có thể tạo cây quá phức tạp không tổng quát hóa dữ liệu tốt. Các cơ chế như cắt tia, thiết lập số lượng mẫu tối thiểu cần thiết tại một nút lá hoặc thiết lập độ sâu tối đa của cây là cần thiết để tránh vấn đề này. Cây quyết định có thể không ổn định vì các biến thể nhỏ trong dữ liệu có thể dẫn đến việc tạo ra một cây hoàn toàn khác. Vấn đề này được giảm thiểu bằng cách sử dụng cây quyết định trong một tập hợp. Các dự đoán của cây quyết định không tron tru cũng không liên tục, mà là các phép gần đúng không đổi từng mảnh như trong hình trên. Do đó, họ không giỏi ngoại suy.

8.2.2. Thuật toán Random Forest (Rừng ngẫu nhiên)

Random Forests là thuật toán học có giám sát, linh hoạt và dễ sử dụng.

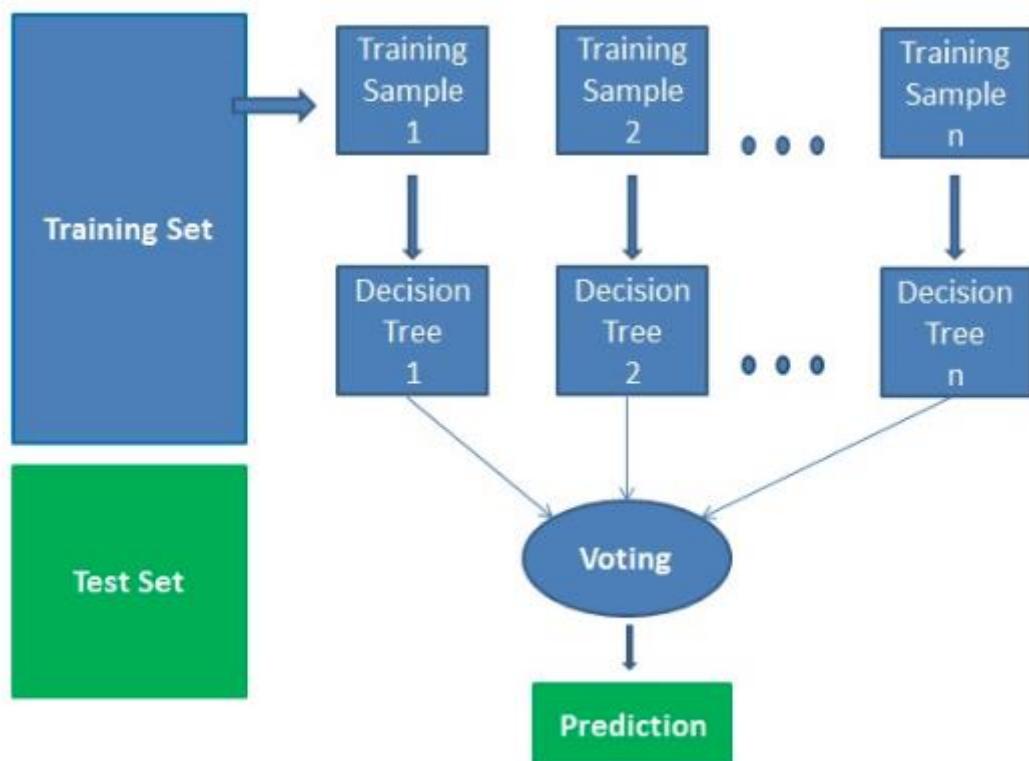
Random forests tạo ra cây quyết định trên các mẫu dữ liệu được chọn ngẫu nhiên, được dự đoán từ mỗi cây và chọn giải pháp tốt nhất bằng cách bỏ phiếu. Nó cũng cung cấp một chỉ báo khá tốt về tầm quan trọng của tính năng. Random forests có nhiều ứng dụng, chẳng hạn như công cụ đề xuất, phân loại hình ảnh và lựa chọn tính năng. Nó có thể được sử dụng để phân loại các ứng viên cho vay trung thành, xác định hoạt động gian lận và dự đoán các bệnh. Nó nằm ở cơ sở của thuật toán Boruta, chọn các tính năng quan trọng trong tập dữ liệu.

Về mặt kỹ thuật, nó là một phương pháp tổng hợp (dựa trên cách tiếp cận phân chia và chinh phục) của các cây quyết định được tạo ra trên một tập dữ liệu được chia ngẫu nhiên. Bộ sưu tập phân loại cây quyết định này còn được gọi là rừng. Cây quyết định riêng lẻ được tạo ra bằng cách sử dụng chỉ báo chọn thuộc tính như tăng thông tin,

tỷ lệ tăng và chỉ số Gini cho từng thuộc tính. Mỗi cây phụ thuộc vào một mẫu ngẫu nhiên độc lập. Trong bài toán phân loại, mỗi phiếu bầu chọn và lớp phổ biến nhất được chọn là kết quả cuối cùng. Trong trường hợp hồi quy, mức trung bình của tất cả các kết quả đầu ra của cây được coi là kết quả cuối cùng. Nó đơn giản và mạnh mẽ hơn so với các thuật toán phân loại phi tuyến tính khác.

Các bước hoạt động của thuật toán Random Forest:

- Chọn các mẫu ngẫu nhiên từ tập dữ liệu đã cho.
- Thiết lập cây quyết định cho từng mẫu và nhận kết quả dự đoán từ mỗi quyết định cây.
- Bỏ phiếu cho mỗi kết quả dự đoán.
- Chọn kết quả được dự đoán nhiều nhất là dự đoán cuối cùng.



Hình 8.2 Các bước hoạt động của thuật toán Random Forest (Nguồn: Random Forests Classifiers, Shivam Sharma)

Ưu điểm: Random forests được coi là một phương pháp chính xác và mạnh mẽ vì số cây quyết định tham gia vào quá trình này. Nó không bị vấn đề overfitting. Lý do chính là nó mất trung bình của tất cả các dự đoán, trong đó hủy bỏ những thành kiến. Thuật toán có thể được sử dụng trong cả hai vấn đề phân loại và hồi quy. Random forests cũng có thể xử lý các giá trị còn thiếu.

Nhược điểm: Random forests chậm tạo dự đoán bởi vì nó có nhiều cây quyết định. Bất cứ khi nào nó đưa ra dự đoán, tất cả các cây trong rừng phải đưa ra dự đoán cho cùng một đầu vào cho trước và sau đó thực hiện bỏ phiếu trên đó. Toàn bộ quá trình này tốn thời gian. Mô hình khó hiểu hơn so với cây quyết định, nơi bạn có thể dễ dàng đưa ra quyết định bằng cách đi theo đường dẫn trong cây.

8.3. Tính cấp thiết trong xây dựng mô hình dự báo học viên tái tục

Học viên tái tục là những học viên có mong muốn tiếp tục thời gian học tập tại trung tâm so với dự kiến ban đầu để nâng cao trình độ.

Số lượng học viên tái tục nhiều phản ánh chất lượng của các khóa học đồng thời phản ánh sự hài lòng của họ. Đây chính là cách marketing hiệu quả, giúp thu hút ngày càng nhiều học viên tới tham gia các khóa học.

Chăm sóc tái tục	<ul style="list-style-type: none"> - Lên danh sách các học viên có kế hoạch tái tục trong tháng. - Triển khai chăm sóc học viên tái tục. - Đưa kế hoạch về doanh số của tháng và chi tiết theo tuần 	<ul style="list-style-type: none"> - Lên danh sách trước mùng 1 hàng tháng - Chăm sóc theo đúng chu trình tái tục 	File tổng <ul style="list-style-type: none"> - Đạt KPI về số lượng chăm sóc tái tục - Đạt KPI về doanh thu tái tục - Đưa rõ lý do về các TH từ chối tái tục
------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Hình 8.3 Chi tiết quy trình chăm sóc học viên tái tục (Nguồn: Công ty Edura)

Chi phí đầu tư cho học viên tái tục ít nhưng tạo ra nguồn doanh thu lớn. Việc tìm ra các học viên tiềm năng là một vấn đề khó trong quá trình vận hành của công ty.

Có rất nhiều yếu tố ảnh hưởng tới quyết định tái tục của học viên, nhưng đâu sẽ là yếu tố quyết định nhiều, đâu là yếu tố ít mang tính quyết định.

Điều này đặt ra yêu cầu cần xây dựng một mô hình dự đoán học viên tái tục để xây dựng danh sách học viên có kế hoạch tái tục giúp tối ưu hóa quy trình vận hành của công ty và không bỏ lỡ các học viên tiềm năng.

Từ những yêu cầu trên, nhóm quyết định xây dựng mô hình “*Dự báo học viên tái tục sử dụng thuật toán rừng ngẫu nhiên và cây quyết định*”. Hai thuật toán này rất hiệu quả trong bài toán hiện tại của công ty, dữ liệu đầu vào sẽ là dữ liệu học viên tái tục trong quá khứ, đã được gán nhãn.

8.4. Xây dựng mô hình dự báo tái tục

8.4.1. Thu thập dữ liệu

Dữ liệu sẽ được lấy từ phòng đào tạo, các cổ văn học tập, giáo viên, chăm sóc viên năm 2021 và tổng hợp thành file dữ liệu thô.

ID_HV	Hoten	Nơi sinh sống	TrìnhDo	Tuổi	ID_KH	Loại Giáo viên	ID_GV	Tháng BD học	Ngayhoc	Phản hồi quá trình chăm sóc	Tái tục
420240761	Phan Thị Loan	Kaiwa 1		29	SKAI31030JP	Nhật Bản	GV175	3	T2-T4-T6	15/03/2022 - BTVN: 0/1 - Nghỉ k phép: 0 - Nghỉ có phép: 0 - Nội dung: HV không gặp vấn đề gì, HV bảo em đi làm nên vẫn chưa làm	Không
420240929	Nguyễn Thị Phương An	Kaiwa 1		17	CKAI121030VN	Việt Nam	GV266	3	T2-T4-T6	19/03/2022: Mạng hơi yếu nghe ít rõ	Không
420240835	Lê thị thu hương	Kaiwa 1		40	CKAI121030VN	Việt Nam	GV221	3	T2-T4-T6	19/03/2022: Chỉ thấy cũng ok chưa thấy cần có gì hỗ trợ cả - Nội dung: HV k gặp vấn đề gì	Có
420240931	-Trương Võ Hoài Linh	Kaiwa 1		43	SKAI21030VN	Việt Nam	GV18	3	T2-T4-T6	19/03/2022: "Chỉ thấy buổi học 30 phút đúng là quá gấp gáp. Buổi kaiwa nên cho	Có
420240760	Nguyễn Hữu Ngọc	Kaiwa 2		22	CKAI121030VN	Việt Nam	GV20	4	T4-T6	17/03/2022 - BTVN: 0/1 - Nghỉ k phép: 0 - Nghỉ có phép: 0 - Nội dung: HV không gặp vấn đề gì, HV đang đi làm nên sẽ hoàn thiện BT	Không
420240841	Nguyễn Đức Minh	Kaiwa 3		18	SKAI21030VN	Việt Nam	GV22	3	T3-T5-T7	19/03/2022: giờ học ngắn	Không
420240815	Phạm Anh Tuấn	PreKaiwa 2		38	CKAI33441030	Việt Nam	GV24	4	T2-T4-T6	Học được 2 buổi HV phản ánh app ứng dụng đơ lõi, ăn thao tác mà không	Không
420240797	Lục Thị Chinh (My)	Kaiwa 3		19	CKAIP211030VN	Việt Nam	GV127	3	T3-T5-T7	19/03/2022: HV chưa biết làm BTVN, đã hướng dẫn HV	Không

Hình 8.4 Dữ liệu thu thập được

Dữ liệu gồm 556 học viên được gán nhãn với các cột phản ánh thông tin như trình độ, ngày học, giáo viên dạy, phản hồi quá trình chăm sóc, ...

Có 62 trên tổng số 556 học viên quyết định tái tục.

8.4.2. Xử lý dữ liệu

Dữ liệu gốc	Dữ liệu sau biến đổi
Từ ngày tháng năm sinh	Độ tuổi
Ngày học	Quy đổi thành 3 loại: 246,357, khác
Giáo viên đang dạy cụ thể	Loại giáo viên
Phản hồi của học viên trong quá trình chăm sóc	Hài lòng, Bình thường, chưa hài lòng
Bài tập về nhà	Chưa làm, Làm ít, Hoàn thành đủ
Đánh giá của CVHT	Tốt, Bình thường, Kém
Đánh giá của GV	Tốt, Bình thường, Kém

Bảng 8.1 Các tiêu chí làm sạch và xử lý dữ liệu

Sau khi làm sạch và xử lý ta thu được bảng dữ liệu như sau:

ID_HV	HoTen	TrinhDo	DoTroi	ID_KH	LoaiGV	ThangBDHoc	NgayHoc	Lam_BTBN	PhanhoiChamSoc	DanhGiaCuaCVHT	DanhGiaCuaGV	Taituc
72019075 Huỳnh Thị Trúc	Kaiwa 2		Trên 30	SKAI21030VN	Viet Nam	10/357		Hoàn thành đủ	Hài lòng	Tốt	Tốt	Có
112019125 Vũ Hoàng Yên	Kaiwa 3		Từ 18-30	SKAI3N060VN	Viet Nam	12/357		Làm ít	Chưa hài lòng	Bình Thường	Bình Thường	Không
112019128 Nguyễn Thị Thai	Kaiwa 3		Từ 18-30	SKAI11030VN	Viet Nam	12/357		Làm ít	Chưa hài lòng	Bình Thường	Kém	Không
112019132 Lê Hoa	Kaiwa 4		Trên 30	SKAI3N060VN	Viet Nam	12/357		Hoàn thành đủ	Bình thường	Bình Thường	Bình Thường	Có
112019140 Nguyễn Tiến Hà	Kaiwa 3		Từ 18-30	SKAI31030VN	Viet Nam	12/246		Chưa làm	Chưa hài lòng	Bình Thường	Bình Thường	Không
420237187 Hoàng Hồng Ánh	Kaiwa 1		Trên 30	SKAI11030VN	Viet Nam	12/246		Chưa làm	Hài lòng	Bình Thường	Bình Thường	Có
420236639 Nguyễn Thị Thai	Kaiwa 2		Tрен 30	SKAI21030VN	Viet Nam	1/246		Làm ít	Bình thường	Bình Thường	Bình Thường	Không
420237074 Ngô Sỹ Trung	Kaiwa 2		Từ 18-30	CKAI231030VN	Viet Nam	12/246		Làm ít	Chưa hài lòng	Bình Thường	Bình Thường	Không
420237698 Nguyễn Út	Kaiwa 2		Tren 30	SKAI21030VN	Viet Nam	12/246		Hoàn thành đủ	Bình thường	Bình Thường	Bình Thường	Không
420237285 Huỳnh Thanh Tí	Kaiwa 2		Từ 18-30	CKAI231030VN	Viet Nam	12/246		Chưa làm	Hài lòng	Tốt	Tốt	Có
420237719 Bùi Văn Quang	Kaiwa 3		Từ 18-30	SKAI121030VN	Viet Nam	1/246		Làm ít	Chưa hài lòng	Bình Thường	Bình Thường	Không
420237758 Nguyễn Thành A	Kaiwa 6		Tren 30	SKAI61030JP	Nhật Bản	12/246		Chưa làm	Bình thường	Bình Thường	Bình Thường	Không
420237699 Khuất Văn Biên	Kaiwa 3		Từ 18-30	SKAI31030JP	Nhật Bản	12/357		Hoàn thành đủ	Bình thường	Tốt	Tốt	Không
420237459 Trần Thị Mỹ Duy	Kaiwa 2		Tren 30	CKAI231030VN	Viet Nam	12/246		Làm ít	Bình thường	Bình Thường	Bình Thường	Không
420237438 Lê Thị Tâm Anh	Kaiwa 2		Tren 30	SKAI21030VN	Viet Nam	1/246		Hoàn thành đủ	Hài lòng	Bình Thường	Bình Thường	Có
420237756 Nguyễn Thị Lưu	Kaiwa 2		Tren 30	CKAI231030VN	Viet Nam	12/246		Làm ít	Bình thường	Tốt	Bình Thường	Không
420237394 Nguyễn Thị Trúc	Kaiwa 1		Từ 18-30	CKAI121030VN	Viet Nam	12/246		Hoàn thành đủ	Chưa hài lòng	Tốt	Bình Thường	Không
420237531 Phạm Quang Sĩ	Kaiwa 2		Từ 18-30	CKAI231030VN	Viet Nam	12/357		Làm ít	Bình thường	Tốt	Bình Thường	Không
420237655 Lê Thị Huyền	Kaiwa 2		Từ 18-30	CKAI231030VN	Viet Nam	1/357		Làm ít	Bình thường	Tốt	Bình Thường	Không
420237812 Dương Thị Hân	Kaiwa 2		Từ 18-30	CKAI231030VN	Viet Nam	12/246		Làm ít	Chưa hài lòng	Tốt	Bình Thường	Không
420237643 Nguyễn Thị Chu	Kaiwa 2		Tren 30	CKAI231030VN	Viet Nam	12/357		Làm ít	Hài lòng	Tốt	Tốt	Có
420237865 Nguyễn Thị Liên	Kaiwa 2		Từ 18-30	CKAI231030VN	Viet Nam	12/357		Chưa làm	Bình thường	Bình Thường	Bình Thường	Không
420237377 Phạm Văn Quy	Kaiwa 2		Tren 30	CKAI231030VN	Viet Nam	1/246		Làm ít	Bình thường	Bình Thường	Bình Thường	Không
420237440 Nguyễn Yến	Kaiwa 1		Từ 18-30	CKAI121030VN	Viet Nam	12/357		Hoàn thành đủ	Chưa hài lòng	Tốt	Bình Thường	Không
420237901 Đinh Văn Hiển	Kaiwa 3		Tren 30	SKAI1030JP	Nhật Bản	12/246		Làm ít	Bình thường	Tốt	Bình Thường	Không
420237748 Lê Thị Ngọc	Kaiwa 3		Dưới 18	CKAI341030JP	Nhật Bản	12/246		Làm ít	Bình thường	Tốt	Bình Thường	Không
420237937 Trần Văn Thành	Kaiwa 3		Từ 18-30	CKAI341030JP	Nhật Bản	12/246		Chưa làm	Bình thường	Bình Thường	Bình Thường	Không
420236955 Trịnh Thị Hồng K	Kaiwa 2		Từ 18-30	CKAI231030VN	Viet Nam	1/246		Hoàn thành đủ	Bình thường	Tốt	Tốt	Không
420237925 Trần Thị Thanh	Kaiwa 2		Từ 18-30	SKAI21030VN	Viet Nam	1/246		Làm ít	Bình thường	Bình Thường	Bình Thường	Không
420238182 Trần Văn Chươi	Kaiwa 2		Tren 30	CKAI231030VN	Viet Nam	12/357		Hoàn thành đủ	Bình thường	Tốt	Tốt	Không
420238103 Vũ Văn Tiến	Kaiwa 1		Tren 30	CKAI121030VN	Viet Nam	12/246		Hoàn thành đủ	Chưa hài lòng	Bình Thường	Bình Thường	Không
420238058 Nguyễn Thị Hậu	Kaiwa 2		Tren 30	CKAI231030VN	Viet Nam	1/246		Làm ít	Bình thường	Bình Thường	Bình Thường	Không
420237839 Phạm Việt Cường	Kaiwa 6		Từ 18-30	SKAI61030JP	Nhật Bản	12/357		Chưa làm	Hài lòng	Bình Thường	Bình Thường	Không

Hình 8.5 Dữ liệu sau khi đã làm sạch

8.4.3. Xây dựng mô hình dự đoán tái tục với hai thuật toán Rừng ngẫu nhiên và Cây quyết định

- Khai báo thư viện:

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn import tree
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.preprocessing import LabelEncoder
from sklearn.tree import plot_tree
from sklearn.metrics import plot_confusion_matrix
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
from sklearn.metrics import classification_report
from tabulate import tabulate
```

- Đọc dữ liệu:

```
[ ] df = pd.read_excel('./datasetHeHoTro.xlsx')

print(df)
```

	ID_HV	HoTen	TrinhDo	DoTuoi	ID_KH
0	72019075	Huỳnh Thị Trúc Ly	Kaiwa 2	Trên 30	SKAI21030VN
1	112019125	Vũ Hoàng Yến	Kaiwa 3	Từ 18-30	SKAI3N060VN
2	112019128	Nguyễn Thị Thanh Kiều	Kaiwa 3	Từ 18-30	SKAI11030VN
3	112019132	Lê Hoa	Kaiwa 4	Trên 30	SKAI3N060VN
4	112019140	Nguyễn Tiến Hải	Kaiwa 3	Từ 18-30	SKAI31030VN
..
560	420229560	Nguyễn Quang Việt	Kaiwa 1	Trên 30	CKAI121030VN
561	420247205	Trương Đinh Tuấn	Kaiwa 1	Trên 30	SKAI11030VN
562	420239068	TRẦN XUÂN THỊNH	Kaiwa 1	Từ 18-30	SKAI1N060VN
563	420245339	Trần Thị Thu	Kaiwa 1	Trên 30	SKAI1N060VN
564	420247377	Tô Văn Thuận	Kaiwa 1	Từ 18-30	CKAI121030VN

	LoaiGV	ThangBDHoc	NgayHoc	Lam_BTVN	PhanhoiChamSoc	\
0	Việt Nam	10	357	Hoàn thành đủ	Hài lòng	
1	Việt Nam	12	357	Làm ít	Chưa hài lòng	
2	Việt Nam	12	357	Làm ít	Chưa hài lòng	
3	Việt Nam	12	357	Hoàn thành đủ	Bình thường	
4	Việt Nam	12	246	Chưa làm	Chưa hài lòng	
..
560	Việt Nam	9	246	Làm ít	Hài lòng	
561	Việt Nam	9	Khác	Hoàn thành đủ	Bình thường	
562	Việt Nam	9	Khác	Làm ít	Hài lòng	
563	Việt Nam	9	Khác	Làm ít	Bình thường	
564	Việt Nam	9	246	Hoàn thành đủ	Bình thường	

- Mô tả dữ liệu

#Số Lượng mẫu và thuộc tính
df.shape

(565, 13)

df.head(10)														
	ID_HV	HoTen	TrinhDo	DoTuoI	ID_KH	LoaiGV	ThangBDHoc	NgayHoc	Lam_BTVN	PhanhoiChamSoc	DanhGiaCuaCVHT	DanhgiaieveGV	Taituc	
0	72019075	Huỳnh Thị Trúc Ly	Kaiwa 2	Trên 30	SKAI21030VN	Việt Nam	10	357	Hoàn thành đủ	Hải lòng	Tốt	Tốt	Có	
1	112019125	Vũ Hoàng Yến	Kaiwa 3	Từ 18-30	SKAI3N060VN	Việt Nam	12	357	Làm ít	Chưa hài lòng	Bình Thường	Bình Thường	Không	
2	112019128	Nguyễn Thị Thanh Kiều	Kaiwa 3	Từ 18-30	SKAI11030VN	Việt Nam	12	357	Làm ít	Chưa hài lòng	Bình Thường	Kém	Không	
3	112019132	Lê Hoa	Kaiwa 4	Trên 30	SKAI3N060VN	Việt Nam	12	357	Hoàn thành đủ	Bình thường	Bình Thường	Bình Thường	Có	
4	112019140	Nguyễn Tiến Hải	Kaiwa 3	Từ 18-30	SKAI3N060VN	Việt Nam	12	246	Chưa làm	Chưa hài lòng	Bình Thường	Bình Thường	Không	
5	420237187	Hoàng Hồng Ánh	Kaiwa 1	Trên 30	SKAI11030VN	Việt Nam	12	246	Chưa làm	Hải lòng	Bình Thường	Bình Thường	Có	
6	420236639	Nguyễn Thị Thanh Tâm	Kaiwa 2	Trên 30	SKAI21030VN	Việt Nam	1	246	Làm ít	Bình thường	Bình Thường	Bình Thường	Không	
7	420237074	Ngô Sỹ Trung	Kaiwa 2	Từ 18-30	CKAI231030VN	Việt Nam	12	246	Làm ít	Chưa hài lòng	Bình Thường	Bình Thường	Không	
8	420237698	Nguyễn Đức Trưởng	Kaiwa 2	Trên 30	SKAI21030VN	Việt Nam	12	246	Hoàn thành đủ	Bình thường	Bình Thường	Bình Thường	Không	
9	420237285	Huynh Thành Tuyền	Kaiwa 2	Từ 18-30	CKAI231030VN	Việt Nam	12	246	Chưa làm	Hải lòng	Tốt	Tốt	Có	

Các giá trị thống kê
df.isnull().sum()

```
ID_HV          0
HoTen          0
TrinhDo        0
DoTuoI         0
ID_KH          0
LoaiGV          0
ThangBDHoc      0
NgayHoc         0
Lam_BTVN        2
PhanhoiChamSoc  2
DanhGiaCuaCVHT  2
DanhgiaieveGV  2
Taituc          2
dtype: int64
```

Hiển thị kiểu dữ liệu của các thuộc tính
df.dtypes

```
ID_HV           int64
HoTen          object
TrinhDo        object
DoTuoI         object
ID_KH          object
LoaiGV          object
ThangBDHoc      int64
NgayHoc         object
Lam_BTVN        object
PhanhoiChamSoc object
DanhGiaCuaCVHT object
DanhgiaieveGV object
Taituc         object
dtype: object
```

- Tiết xử lý dữ liệu

Đổi dữ liệu từ dạng định danh (object) về dạng số

	ID_HV	HoTen	TrinhDo	DoTuoi	ID_KH	LoaiGV	ThangBDHoc	NgayHoc	Lam_BTVN	PhanhoiChamSoc	DanhGiaCuaCVHT	DanhgiaieveGV	Taituc
0	72019075	60	1	1	18	1	10	1	1	2	2	2	0
1	112019125	486	2	2	21	1	12	1	2	1	0	0	1
2	112019128	247	2	2	16	1	12	1	2	1	0	1	1
3	112019132	85	3	1	21	1	12	1	1	0	0	0	0
4	112019140	271	2	2	20	1	12	0	0	1	0	0	1
5	420237187	34	0	1	16	1	12	0	0	2	0	0	0
6	420236639	249	1	1	18	1	1	0	2	0	0	0	1
7	420237074	307	1	2	6	1	12	0	2	1	0	0	1
8	420237698	303	1	1	18	1	12	0	1	0	0	0	1
9	420237285	50	1	2	6	1	12	0	0	2	2	2	0

Chuẩn bị tập dữ liệu huấn luyện và test (trong đó, dữ liệu huấn luyện – 70%, dữ liệu test – 30%) đồng thời bỏ hai cột ID_HV và HoTen

5.1. Chuẩn bị tập dữ liệu huấn luyện (train) và tập dữ liệu kiểm thử (test)

```
# Xác định các thuộc tính độc lập X và thuộc tính phụ thuộc y
features = ['TrinhDo', 'DoTuoi', 'ID_KH', 'LoaiGV', 'ThangBDHoc', 'NgayHoc', 'Lam_BTVN', 'PhanhoiChamSoc', 'DanhGiaCuaCVHT', 'DanhgiaieveGV']

target = ['Taituc']
X = df[features]
y = df[target]

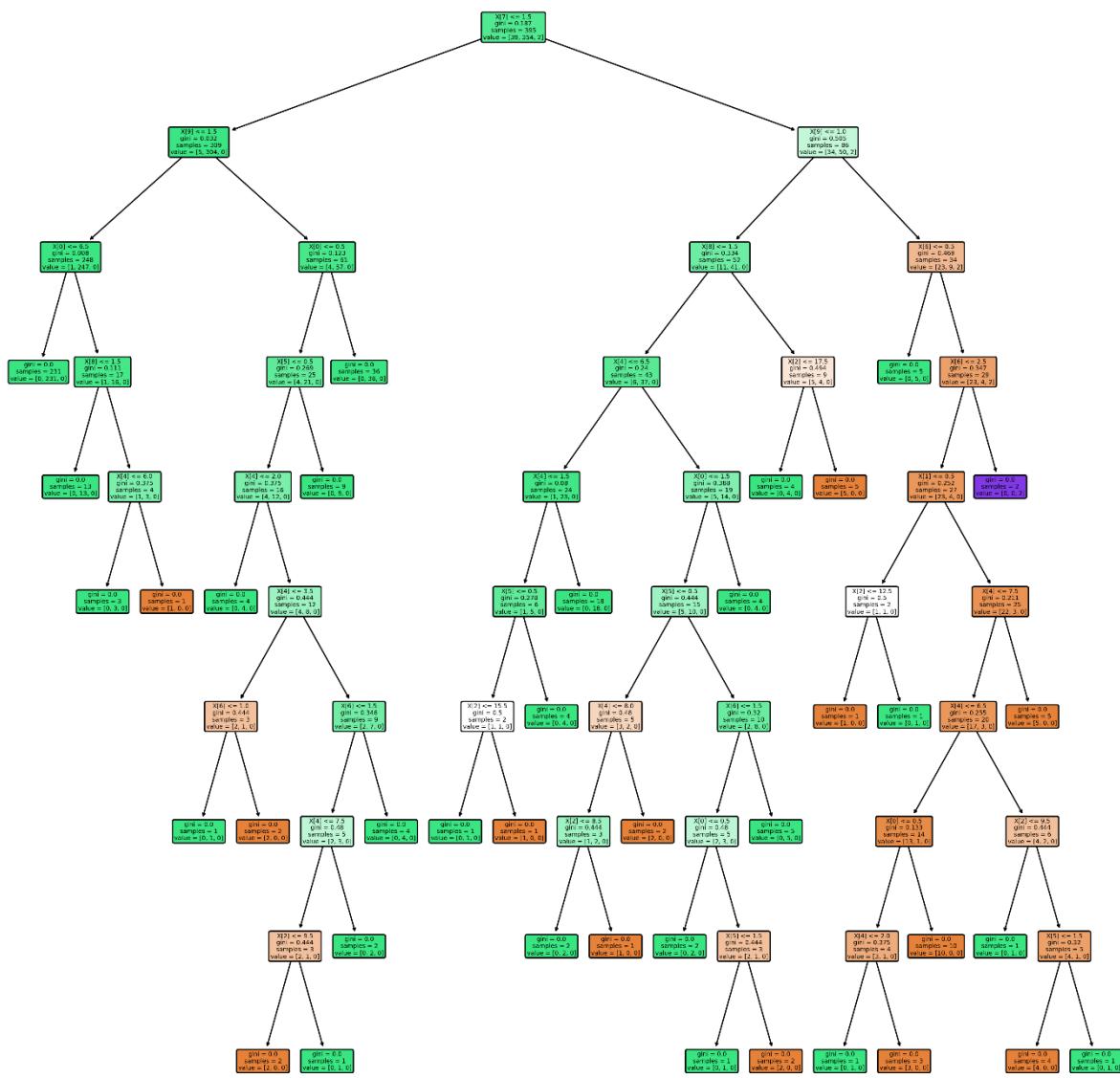
# Chia bộ dữ liệu thành hai tập: train & test (theo tỉ lệ 70% & 30%)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3)
```

8.4.3.1. Mô hình Decision Tree

```
# Lựa chọn kỹ thuật học cây quyết định (thêm các tham số nếu cần can thiệp sâu hơn)
model = tree.DecisionTreeClassifier()

# Đưa dữ liệu vào huấn luyện mô hình
model = model.fit(X_train, y_train)
```

Cây quyết định sau khi chạy mô hình:



Hình 8.6 Hình ảnh cây quyết định sau khi chạy mô hình Decision Tree

8.4.3.2. Thuật toán Random Forest

```
[ ] # Lựa chọn kỹ thuật học Random Forest (thêm các tham số nếu cần can thiệp sâu hơn)
modelRandomForest = RandomForestClassifier()

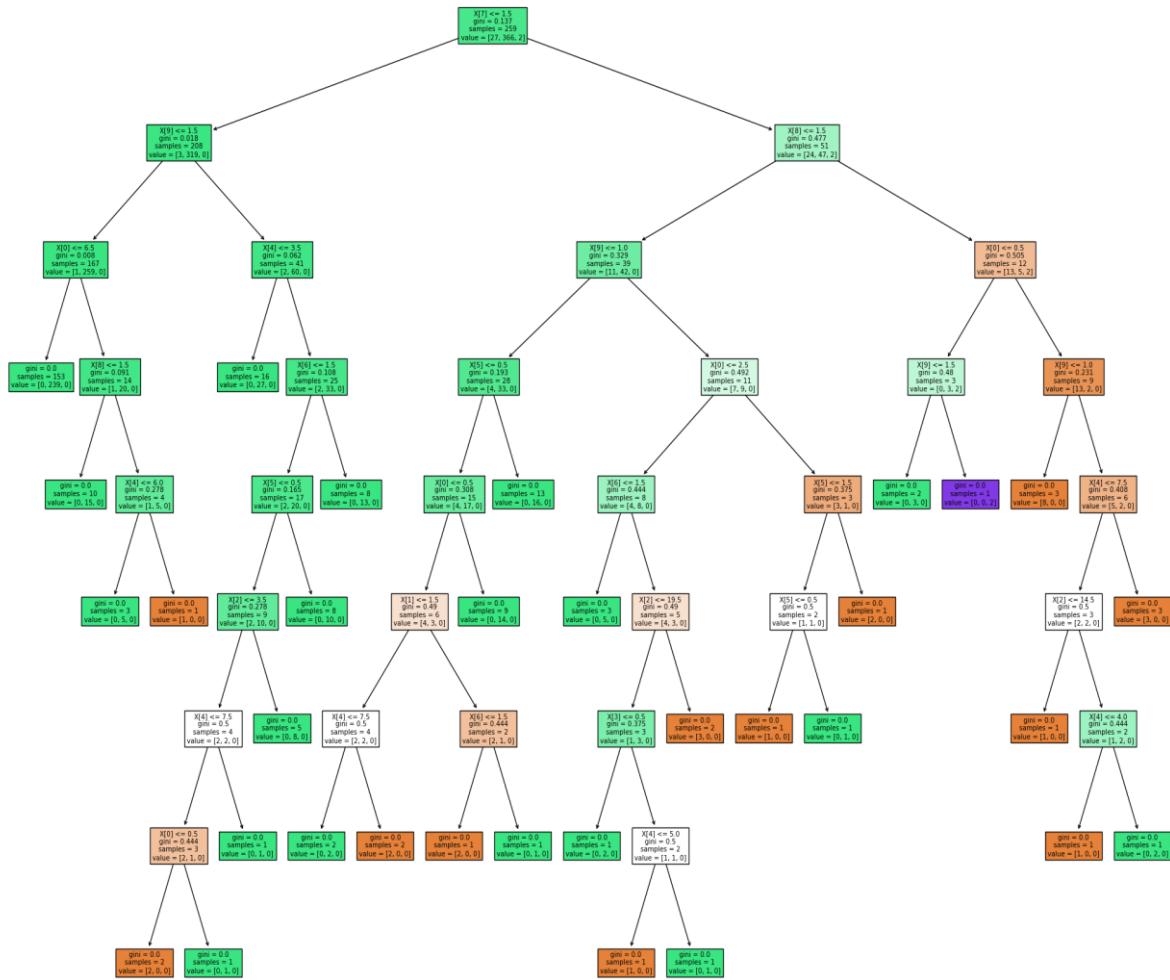
# Đưa dữ liệu vào huấn luyện mô hình
modelRandomForest.fit(X_train, y_train)
```

Hiển thị một cây trong rừng ngẫu nhiên

```
[ ] # Hiển thị mô hình: ví dụ một cây trong rừng ngẫu nhiên
estimator = modelRandomForest.estimators_[1] # cây thứ 2 của rừng
plot_tree(estimator, filled = True)
plt.show()

# Lưu lại cây dưới dạng ảnh nếu muốn
figRandomForest = plt.figure(figsize = (25,20))
_= plot_tree(estimator, filled = True)
figRandomForest.savefig("random_Forest.png")
```

Cây quyết định sau khi chạy mô hình:



Hình 8.7 Hình ảnh cây quyết định sau khi chạy thuật toán Random Forest

8.5. Kiểm thử mô hình dự đoán với Decision Tree và Random Forest

Với dữ liệu chạy thử để dự đoán:

Trình độ	Kaiwa 2	1
Độ tuổi	Trên 30	1
ID_KH	SKAI21030VN	18
Loại GV	Việt Nam	1
Tháng bắt đầu học	10	10
Ngày học	357	1
Làm bài tập về nhà	Hoàn thành đủ	1
Phản hồi chăm sóc	Hài lòng	2
Đánh giá của CVHT	Tốt	2
Đánh giá về GV	Tốt	2

Bảng 8.2 Mô tả dữ liệu đầu vào cho quá trình kiểm thử mô hình dự đoán

```
# Sử dụng mô hình dự đoán dự đoán tái tiếp tục đăng ký khóa học của học viên sau khi
# Trình độ 1, Độ tuổi 1, ID_KH 18, Loại GV 1, Tháng bắt đầu học 10, Ngày học 1, Làm
new_hocvien = [[1,1,18,1,10,1,1,2,2,2]]
predicted_label = model.predict(new_hocvien)
if predicted_label == 1:
    print("Khách hàng có khả năng tái tục đăng ký khóa học (Desision tree)")
else:
    print("Khách hàng không có khả năng tái tục đăng ký khóa học (Desision tree)")
```

Khách hàng không có khả năng tái tục đăng ký khóa học (Desision tree)

Hình 8.8 Kết quả chạy mô hình dự đoán với mô hình Decision Tree

```
# Sử dụng mô hình dự đoán dự đoán tái tiếp tục đăng ký khóa học của học viên sau khi
# Trình độ 1, Độ tuổi 1, ID_KH 18, Loại GV 1, Tháng bắt đầu học 10, Ngày học 1, Làm b
new_hocvien_Random_Forest = [[1,1,18,1,10,1,1,2,2,2]]
predicted_label = modelRandomForest.predict(new_hocvien_Random_Forest)
if predicted_label == 1:
    print("Khách hàng có khả năng tái tục đăng ký khóa học (Random Forest)")
else:
    print("Khách hàng không có khả năng tái tục đăng ký khóa học (Random Forest)")
```

Khách hàng không có khả năng tái tục đăng ký khóa học (Random Forest)

Hình 8.9 Kết quả chạy mô hình dự đoán với mô hình Random Forest

8.6. Kiểm tra mô hình và đánh giá độ chính xác của mô hình

8.6.1. Độ chính xác của mô hình

```
# Kiểm thử mô hình trên tập test
y_pred_array = []
for i in range(1000):
    y_pred_array.append(model.predict(X_test.values))
y_pred = sum(y_pred_array) / len(y_pred_array)
print ("Độ chính xác khi sử dụng decision tree (Trung bình):", accuracy_score(y_test, y_pred) * 100)

# Kiểm thử mô hình trên tập test
y_pred_RandomForest_array = []
for i in range(1000):
    y_pred_RandomForest_array.append(modelRandomForest.predict(X_test.values))
y_pred_RandomForest = sum(y_pred_RandomForest_array) / len(y_pred_RandomForest_array)
print ("Độ chính xác khi sử dụng Random Forest (Trung bình):", accuracy_score(y_test, y_pred_RandomForest) * 100)
```

Độ chính xác khi sử dụng decision tree (Trung bình): 92.3529411764706

Độ chính xác khi sử dụng Random Forest (Trung bình): 93.52941176470588

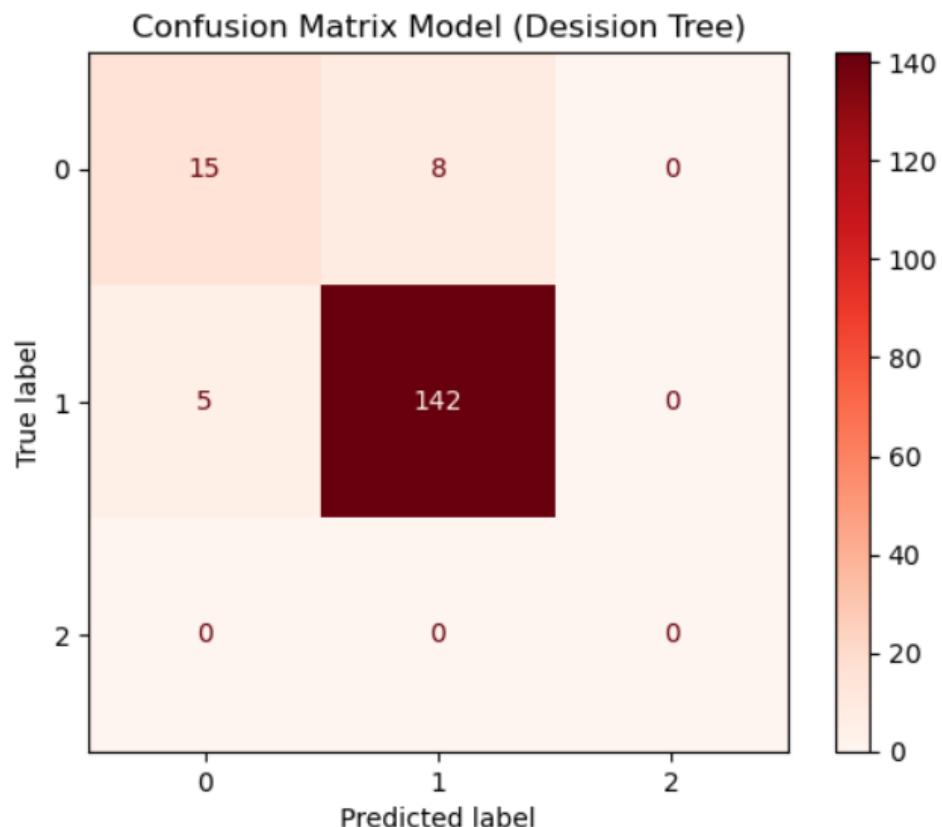
Từ kết quả trên ta có bảng so sánh dưới đây:

	Decision Tree	Random Forest
Độ chính xác	92.35%	93.52%

Bảng 8.3 So sánh độ chính xác giữa Decision Tree và Random Forest

8.6.2. Ma trận nhầm lẫn

- Decision Tree



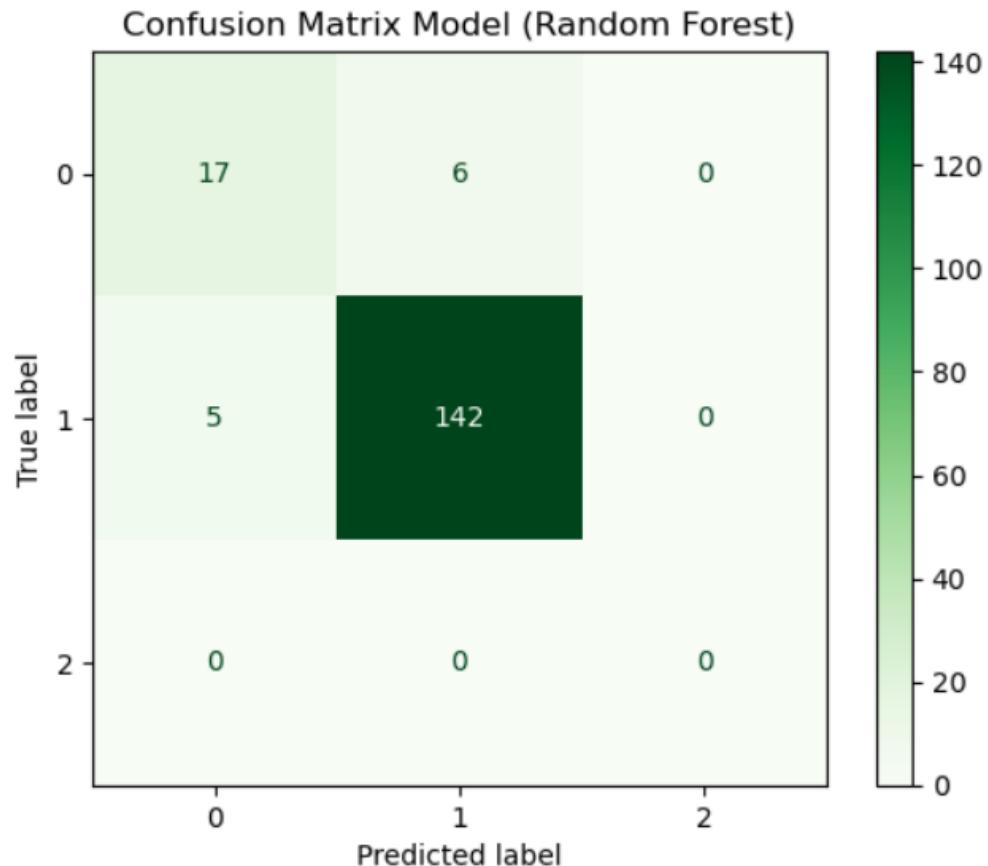
Hình 8.10 Ma trận nhầm lẫn mô hình Decision Tree

```
#Decision Tree
reprot = classification_report(y_test, y_pred)
print(reprot)
```

	precision	recall	f1-score	support
0	0.75	0.65	0.70	23
1	0.95	0.97	0.96	147
accuracy			0.92	170
macro avg	0.85	0.81	0.83	170
weighted avg	0.92	0.92	0.92	170

Hình 8.11 Kết quả độ đo đánh giá của bài toán với mô hình Decision Tree

- Random Forest



Hình 8.12 Ma trận nhầm lẫn thuật toán Random Forest

```
#Random Forest
reprotRandomForest = classification_report(y_test, y_pred_RandomForest)
print(reprotRandomForest)

precision    recall   f1-score   support
      0       0.77     0.74     0.76      23
      1       0.96     0.97     0.96     147
accuracy                           0.94     170
macro avg       0.87     0.85     0.86     170
weighted avg     0.93     0.94     0.93     170
```

Hình 8.13 Kết quả độ đo đánh giá của bài toán với mô hình Random Forest

```
#Confusion Matrix
tn, fp, fn, tp = confusion_matrix(y_test, y_pred).ravel()
tnRF, fpRF, fnRF, tpRF = confusion_matrix(y_test, y_pred_RandomForest).ravel()
data = [['Decision Tree', tn, fp, fn, tp], ['Random Forest', tnRF, fpRF, fnRF, tpRF]]
col_names = ['Model', 'TN', 'FP', 'FN', 'TP']
print(tabulate(data, headers=col_names))
```

Model	TN	FP	FN	TP
Decision Tree	15	8	5	142
Random Forest	17	6	5	142

Hình 8.14 So sánh ma trận nhầm lẫn giữa Decision Tree và Random Forest

Kết luận: Kết quả $precision = \frac{TP}{TP + FP} = \frac{15}{15 + 5} = 0,75$ (Decision Tree) và $precision = \frac{TP}{TP + FP} = \frac{17}{17 + 5} = 0,77$ (Random Forest) khá cao. Như vậy, có thể thấy Precision của mô hình Random Forest có độ chính xác của các điểm tìm được cao hơn so với mô hình Decision Tree. Precision sẽ cần được coi trọng hơn khi lựa chọn model với các bài toán cụ thể khi mà việc nhận nhầm False Positive mang lại kết quả tồi tệ.

$Recall = \frac{TP}{TP + FN} = \frac{15}{15 + 8} = 0,65$ (Decision Tree) và $Recall = \frac{TP}{TP + FN} = \frac{17}{17 + 6} = 0,74$ do đó ta thấy tỷ lệ giữa các điểm positive thực được nhận đúng trên tổng điểm positive thực và kết quả tỷ lệ bỏ sót các sample positive thực khi ứng dụng mô hình Decision Tree cao hơn khi ứng dụng mô hình Random Forest. Do đó, Recall nên được gán trọng số cao hơn khi cân nhắc lựa chọn model tốt nhất khi mà việc nhận nhầm các nhãn Positive thực thành False Negative mang lại hậu quả khôn lường.

Từ kết quả ma trận nhầm lẫn của 2 mô hình trên ta thấy rằng F1 của Decision Tree thấp hơn với mô hình Random Forest nên để lựa chọn mô hình phù hợp với bài toán này thì chúng ta sẽ lựa chọn mô hình của Random Forest vì khi đó chỉ số dung hòa giữa Recall và Precision là yếu tố quyết định giúp chúng ta có căn cứ để lựa chọn model.

Với độ chính xác khi chạy 2 thuật toán là Random Forest và Decision Tree, có thấy thể thấy rằng độ chính xác khi dự đoán lên tới trên 90%, trong tương lai gần nhóm sẽ tìm cách cải thiện độ chính xác gần như tuyệt đối để hỗ trợ tối đa trong việc dự báo khách hành tái tiếp tục đăng ký khóa học tại trung tâm.

KẾT LUẬN

Doanh nghiệp cần phải đưa ra những quyết định kịp thời và đúng đắn thì mới có thể đạt tới thành công. Quyết định doanh nghiệp cũng phụ thuộc vào các yếu tố bên ngoài như tình hình kinh tế, hoàn cảnh thị trường, chiến lược của đối thủ cạnh tranh, vận hành của công ty, và thay đổi của công nghệ... Do đó, những người quản lý ra quyết định phải có mọi thông tin cần thiết. Nhiều công ty cần khai thác triệt để các nguồn dữ liệu để đưa ra các phân tích nhằm hướng tới phương hướng phát triển. Đó là lí do tại sao các công cụ công nghệ thông tin (CNTT) như hệ trợ giúp quyết định (BI) đã được phát triển để giúp cho người quản lý ra quyết định tốt hơn.

Công cụ BI có khả năng thu thập mọi dữ liệu liên quan, phân tích và tổ chức chúng dưới dạng hiển thị, dễ dàng cho người quản lý kiểm điểm rồi ra quyết định. Trong quá khứ dữ liệu được thu thập bởi quan sát và làm tài liệu nhưng khi dữ liệu được thu thập, vẫn có những lỗi phạm phải. Dữ liệu không liên quan hay sai có thể đưa tới lỗ l—————————————————————n và quyết định kém. Để tránh điều đó, phần lớn các công ty đang dựa trên công nghệ thông tin như kinh doanh thông minh (BI) thay vì để con người thu thập và phân tích dữ liệu. Dữ liệu được thu thập một cách tự động bởi hệ thống CNTT và được phân tích bởi phần mềm đặc biệt, bất kì cái gì móc nối với vấn đề đều được thể hiện trong các báo cáo và được gửi tới cấp quản lý để ra quyết định. BI là hệ thống phần mềm tương tác làm việc thu thập mọi thông tin liên quan từ nhiều nguồn như vận hành, thị trường, thu nhập, chi phí, xu hướng, và mô hình doanh nghiệp. Những dữ liệu này được lưu trong các cơ sở dữ liệu nơi phần mềm khai phá dữ liệu có thể duyệt tìm thông tin liên quan và tổ chức chúng thành các báo cáo cho người quản lý.

Áp dụng Machine Learning, AI, những thuật toán học sâu, phân cụm, phân lớp để hiểu rõ dữ liệu của mình hiện có cũng như dự đoán được điều gì sẽ xảy ra. Dựa vào tình hình kinh doanh ngày càng được mở rộng, dữ liệu ngày càng lớn, nhiều loại khác nhau (có cấu trúc, phi cấu trúc), nhiều nguồn thu thập từ những hệ thống của công ty nên việc xây dựng Data Lake cũng là cần thiết.

Với nguồn dữ liệu của công ty ETDE thu thập được thì nhóm chúng em đã sử dụng được công cụ BI, những thuật toán học sâu, phân cụm để có thể lập ra các bảng báo cáo, đưa ra các tình hình của công ty để từ đó có hướng đi và mục tiêu phát triển, các chiến lược kinh doanh phù hợp.

Do sự hạn chế về thời gian, bài báo cáo vẫn còn một số hạn chế nên rất mong nhận được sự đóng góp và nhận xét từ cô để bài báo cáo được hoàn thiện hơn.

TÀI LIỆU THAM KHẢO

- Anon., n.d. *AWS Documentation*. [Online]
 Available at: <https://docs.aws.amazon.com/index.html>
- Anon., n.d. *Data Lake on AWS*. [Online]
 Available at: <https://aws.amazon.com/solutions/implementations/data-lake-solution/>
- Anon., n.d. *Entity Identification Problem in Data Mining*. [Online]
 Available at: <https://www.javatpoint.com/entity-identification-problem-in-data-mining>
- Anon., n.d. *What is a data lake?*. [Online]
 Available at: <https://aws.amazon.com/big-data/what-is-a-data-lake/>
- Center, O. H., n.d. *Understanding Slowly Changing Dimensions*. [Online]
 Available at:
https://docs.oracle.com/cd/E41507_01/epm91pbr3/eng/epm/phcw/concept_UnderstandingSlowlyChangingDimensions-405719.html
- Chen, Z., 2001. *Intelligent Data Warehousing: From Data Preparation to Data Mining*. s.l.:CRC Press.
- Haroon, M., 2020. *Multiple Source Data Processing and Integration*. [Online]
 Available at: <https://datascience.foundation/datatalk/multiple-source-data-processing-and-integration>
- IBM Cloud Education, 2020. *What is Machine Learning?*. [Online]
 Available at: <https://www.ibm.com/cloud/learn/machine-learning>
- O'Leary, D., n.d. REAL-D: A Schema for Data Warehouses. Volume 13, pp. 49-62.
- Ross, R. K. & M., 2013. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. 3rd ed. s.l.:John Wiley & Sons, Inc.
- Shaikh, M. M., 2013. *Create and Populate Date Dimension for Data Warehouse*. [Online]
 Available at: <https://www.codeproject.com/Articles/647950/Create-and-Populate-Date-Dimension-for-Data-Wareho>

Tableau Software, 2022. *Tableau Desktop and Web.* [Online]
Available at: <https://www.tableau.com/support/help>

Taylor, D., 2022. *What is Data Lake? It's Architecture: Data Lake Tutorial.* [Online]

Available at: <https://www.guru99.com/data-lake-architecture.html>

T, N., 2020. *What is Data Integration? Definition, Issues, Techniques and Tools.* [Online]

Available at: <https://binaryterms.com/data-integration.html>

PHỤ LỤC

Phụ lục 1: Nguồn dữ liệu mà nhóm tổng hợp được từ công ty

https://docs.google.com/spreadsheets/d/15tqrB7c0veiOg8REis7kf8iAxmMYVLG2DP_hlBnGg4o/edit#gid=2097753681

- Thông tin dữ liệu học:

A	B	C	D	E	F	G	H	I	J	K
ID_HV	ID_KH	ID_GV	Trangthailop	NgayBD	NgayKT_BL	Ngayhoc	Tongbobuoi	Sobuoiconlai	Sobuoidahoc	
72019001	SKAI1N060VN	GV89	Kết thúc	2019-07-24	2019-11-11	T3-T5-T7	30	0	30	
72019002	SKAI3N060VN	GV47	Kết thúc	2019-07-24	2019-11-11	T3-T5-T7	30	0	30	
72019003	SKAI11030VN	GV104	Kết thúc	2019-07-27	2019-11-14	T3-T5-T7	30	0	30	
72019004	SKAI31030VN	GV105	Kết thúc	2019-07-29	2019-11-16	T3-T5-T7	30	0	30	
72019005	SKAI51030JP	GV01	Kết thúc	2019-07-30	2019-11-17	T3-T5-T7	30	0	30	
72019006	SKAI2N060VN	GV106	Kết thúc	2019-07-31	2019-11-18	T3-T5-T7	30	0	30	
72019007	SKAI41030JP	GV02	Kết thúc	2019-07-31	2019-11-18	T2-T4-T6	30	0	30	
72019008	SKAI51030JP	GV03	Kết thúc	2019-08-03	2019-11-21	T3-T5-T7	30	0	30	
72019009	SKAI11030VN	GV127	Kết thúc	2019-08-04	2019-11-22	T3-T5-T7	30	0	30	
72019010	SKAI31030VN	GV129	Kết thúc	2019-08-05	2019-11-23	T3-T5-T7	30	0	30	
72019011	SKAI31030VN	GV130	Kết thúc	2019-08-05	2019-11-23	T3-T5-T7	30	0	30	
72019012	SKAI21030VN	GV135	Kết thúc	2019-08-07	2019-11-25	T2-T4-T6	30	0	30	
72019013	SKAI31030VN	GV148	Kết thúc	2019-08-07	2019-11-25	T3-T5-T7	30	0	30	
72019014	SKAI11030VN	GV149	Kết thúc	2019-08-10	2019-11-28	T3-T5-T7	30	0	30	
72019015	SKAI21030VN	GV152	Kết thúc	2019-08-10	2019-11-28	T3-T5-T7	30	0	30	
72019016	SKAI11030VN	GV173	Kết thúc	2019-08-11	2019-11-29	T3-T5-T7	30	0	30	
72019017	SKAI11030VN	GV174	Kết thúc	2019-08-12	2019-11-30	T3-T5-T7	30	0	30	
72019018	SKAI11030VN	GV176	Kết thúc	2019-08-13	2019-12-01	T3-T5-T7	30	0	30	
72019019	SKAI11030VN	GV188	Kết thúc	2019-08-13	2019-12-01	T2-T4-T6	30	0	30	
72019020	SKAI11030VN	GV189	Kết thúc	2019-08-13	2019-12-01	T3-T5-T7	30	0	30	
72019021	CKAI121030VN	GV190	Kết thúc	2019-08-14	2020-03-26	T3-T5-T7	60	0	60	
72019022	SKAI11030VN	GV207	Kết thúc	2019-08-14	2019-12-02	T2-T4-T6	30	0	30	
72019023	SKAI21030VN	GV208	Kết thúc	2019-08-14	2019-12-02	T2-T4-T6	30	0	30	
72019024	SKAI11030VN	GV209	Kết thúc	2019-08-15	2019-12-03	T2-T4-T6	30	0	30	
72019025	SKAI11030VN	GV211	Kết thúc	2019-08-14	2019-12-02	T2-T4-T6	30	0	30	
72019026	SKAI11030VN	GV225	Kết thúc	2019-08-20	2019-12-08	T3-T5-T7	30	0	30	
72019027	SKAI11030VN	GV227	Kết thúc	2019-08-16	2019-12-04	T3-T5-T7	30	0	30	

- Thông tin dữ liệu thanh toán:

ID_HV	ID_KH	TrangthaiTT	NgayTT	Nganhnganh (KT)	Sotiennop_Nhat	Sotiennop_Viet	Nganhnganh (sale)	
72019001	SKAI1N060VN	Đặt cọc	2019-07-21	BIDV (21610000537698)		1000000	BIDV	
72019002	SKAI3N060VN	Đặt cọc	2019-07-21	BIDV (21610000537698)		1000000	BIDV	
72019003	SKAI11030VN	Đặt cọc	2019-07-24	BIDV (21610000537698)		1000000	BIDV	
72019004	SKAI31030VN	Đặt cọc	2019-07-26	BIDV (21610000537698)		1000000	BIDV	
72019005	SKAI51030JP	Đặt cọc	2019-07-27	BIDV (21610000537698)		1000000	BIDV	
72019006	SKAI2N060VN	Đặt cọc	2019-07-28	BIDV (21610000537698)		1000000	BIDV	
72019007	SKAI41030JP	Đặt cọc	2019-07-28	BIDV (21610000537698)		1000000	BIDV	
72019008	SKAI51030JP	Đặt cọc	2019-07-31	BIDV (21610000537698)		1000000	BIDV	
72019009	SKAI11030VN	Đặt cọc	2019-08-01	BIDV (21610000537698)		1000000	BIDV	
72019010	SKAI31030VN	Đặt cọc	2019-08-02	BIDV (21610000537698)		1400000	BIDV	
72019011	SKAI31030VN	Đặt cọc	2019-08-02	BIDV (21610000537698)		1000000	BIDV	
72019012	SKAI21030VN	Đặt cọc	2019-08-04	BIDV (21610000537698)		1000000	BIDV	
72019013	SKAI31030VN	Đặt cọc	2019-08-04	VPBank (165191757)		1000000	VP Bank	
72019014	SKAI11030VN	Đặt cọc	2019-08-07	BIDV (21610000537698)		1000000	BIDV	
72019015	SKAI21030VN	Đặt cọc	2019-08-07	BIDV (21610000537698)		1000000	BIDV	
72019016	SKAI11030VN	Đặt cọc	2019-08-08	BIDV (21610000537698)		1000000	BIDV	
72019017	SKAI11030VN	Đặt cọc	2019-08-09	BIDV (21610000537698)		1000000	BIDV	
72019018	SKAI11030VN	Đặt cọc	2019-08-10	BIDV (21610000537698)		1000000	BIDV	
72019019	SKAI11030VN	Đặt cọc	2019-08-10	BIDV (21610000537698)		1000000	BIDV	
72019020	SKAI11030VN	Đặt cọc	2019-08-10	Nhật 1 (Mai - 10160-688566	5000	1075000	Ngân hàng bưu điện Nhật Bản	
72019021	CKAI121030VN	Đặt cọc	2019-08-11	BIDV (21610000537698)		1000000	BIDV	
72019022	SKAI11030VN	Đặt cọc	2019-08-11	BIDV (21610000537698)		1000000	BIDV	
72019023	SKAI21030VN	Đặt cọc	2019-08-11	BIDV (21610000537698)		1000000	BIDV	
72019024	SKAI11030VN	Đặt cọc	2019-08-11	BIDV (21610000537698)		1000000	BIDV	
72019025	SKAI11030VN	Đặt cọc	2019-08-12	BIDV (21610000537698)		1000000	BIDV	
72019026	SKAI11030VN	Đặt cọc	2019-08-18	BIDV (21610000537698)		1000000	BIDV	
72019027	SKAI11030VN	Đặt cọc	2019-08-14	VPBANK CÔNG TY (228504866)		500000	VPBank công ty	

- Thông tin dữ liệu đăng ký:

	A	B	C	D	E	F	G
1	ID_HV	ID_NV	ID_KH	Ngaydk			
2	72019001	Trinhht	SKAI1N060VN	2019-07-19			
3	72019002	Trinhht	SKAI3N060VN	2019-07-19			
4	72019003	Dinhtc	SKAI11030VN	2019-07-22			
5	72019004	Dinhtc	SKAI31030VN	2019-07-24			
6	72019005	Dinhtc	SKAI51030JP	2019-07-25			
7	72019006	Dinhtc	SKAI2N060VN	2019-07-26			
8	72019007	Dinhtc	SKAI41030JP	2019-07-26			
9	72019008	Dinhtc	SKAI51030JP	2019-07-29			
10	72019009	Dinhtc	SKAI11030VN	2019-07-30			
11	72019010	Dinhtc	SKAI31030VN	2019-07-31			
12	72019011	Dinhtc	SKAI31030VN	2019-07-31			
13	72019012	Dinhtc	SKAI21030VN	2019-08-02			
14	72019013	Dinhtc	SKAI31030VN	2019-08-02			
15	72019014	Dinhtc	SKAI11030VN	2019-08-05			
16	72019015	Dinhtc	SKAI21030VN	2019-08-05			
17	72019016	Dinhtc	SKAI11030VN	2019-08-06			
18	72019017	huyenlm	SKAI11030VN	2019-08-07			
19	72019018	Dinhtc	SKAI11030VN	2019-08-08			
20	72019019	huyenlm	SKAI11030VN	2019-08-08			
21	72019020	huyenlm	SKAI11030VN	2019-08-08			
22	72019021	Dinhtc	CKAI121030VN	2019-08-09			
23	72019022	Dinhtc	SKAI11030VN	2019-08-09			
24	72019023	Dinhtc	SKAI21030VN	2019-08-09			
25	72019024	Dinhtc	SKAI11030VN	2019-08-10			
26	72019025	huyenlm	SKAI11030VN	2019-08-09			
27	72019026	huyenlm	SKAI11030VN	2019-08-15			
28	72019027	huyenlm	SKAI11030VN	2019-08-11			

☰ HocVien ▾ HocVien_Salw ▾ KhoaHoc ▾ GiaoVien ▾ NVSale ▾ DangKy ▾

- Thông tin dữ liệu nhân viên sale:

A	B	C	D	E	F
MaNSale	Hoten	Gioitinh			
Trinhhht	Hoàng Thị Trinh	Nữ			
Dinhhtc	Trần Công Định	Nam			
huyenlm	Lê Minh Huyền	Nữ			
Tridm	Đặng Minh Trí	Nam			
Dungtt1	Trần Thị Dung	Nữ			
Linhntm	Nguyễn Thị Mỹ Linh	Nữ			
Quynhln	Lý Ngọc Quỳnh	Nữ			
Anhvtl	Vũ Thị Lan Anh	Nữ			
Thuantv	Trần Văn Thuận	Nam			
Khuent	Nguyễn Thanh Khuê	Nữ			
Ninhnht	Nguyễn Thị Hoài Ninh	Nữ			
Trangdtl	Đặng Thị Lam Trang	Nữ			
Thaott	Trần Thị Thảo	Nữ			
PDT	Phòng Đào Tạo	Nữ			
Minhha	Hà Ánh Minh	Nữ			
Huongnt	Nguyễn Thị Hương	Nữ			
Huongnt1	Nguyễn Thu Hương	Nữ			
Hoaidtt	Đặng Thị Thanh Hoài	Nữ			
Longph	Phan Hoàng Long	Nam			
Linhtk	Trần Khánh Linh	Nữ			
Huongvtl	Vũ Thị Lan Hương	Nữ			
Oanhhtk	Trương Thị Khanh Oanh	Nữ			
Daonta	Nguyễn Thị Ánh Đào	Nữ			
Huyenptk	Phan Thị Khanh Huyền	Nữ			
Nguyenln	Lê Thị Ngọc Nguyên	Nữ			
Linhtna	Trần Nguyễn Ánh Linh	Nữ			
Lanttt	Trần Thị Thu Lan	Nữ			



HocVien ▾

HocVien_Salw ▾

KhoaHoc ▾

GiaoVien ▾

NVSale ▾

- Thông tin dữ liệu giáo viên:

A	B	C	D	E	F	G
ID_GV	HoTen	QuocTich	Email	Gioitinh	Mucluong	
GV01	Ichinose Miho	Nhật Bản	iamicns01@gmail.com	Nữ	950	
GV02	Nakao Yuta	Nhật Bản	sakuratoso.nakao@gmail.com	Nữ	950	
GV03	Chiharu Yabuki	Nhật Bản	chiharu458@gmail.com	Nữ	950	
GV04	Phan Nhung	Việt Nam	Phannhungv97@gmail.com	Nữ	120000	
GV05	Kamiunten Hiroko	Nhật Bản	heroxxkamiunten@gmail.com	Nữ	950	
GV06	Miyata Mariko	Nhật Bản	t.mariko.shun@gmail.com	Nữ	950	
GV07	Đinh Văn Hoàng	Việt Nam	hoangdinh.tohoku@gmail.com	Nam	150000	
GV08	Đặng Thị Thu Hà	Việt Nam	dangha94bg@gmail.com	Nữ	120000	
GV09	Võ Chí Thiện	Việt Nam	chithien1993spk@gmail.com	Nam	120000	
GV10	Kinugasa Moe	Nhật Bản	pro.kinugasa@gmail.com	Nam	950	
GV12	Aya Miura	Nhật Bản	ayamanma1019@docomo.ne.jp	Nữ	950	
GV13	Shinohe Yuto	Nhật Bản	nicc.bb.nb.510@icloud.com	Nam	950	
GV14	Lê Thị Lai	Việt Nam	lethilai757@gmail.com	Nữ	120000	
GV99	Endou yuuta	Nhật Bản	atuy117@gmail.com	Nữ	950	
GV15	Bandou Mana	Nhật Bản	mn.bnd0@gmail.com	Nữ	950	
GV31	Matsuda takashi	Nhật Bản	takahashi3824@gmail.com	Nữ	950	
GV16	Okuda kazuya	Việt Nam	okuokuoku999@gmail.com	Nữ	950	
GV17	Hoàng Thị Xuân Quỳnh	Việt Nam	xuanquynh97.hlu@gmail.com	Nữ	130000	
GV18	Nguyễn Thu Hà	Việt Nam	nguyenthuhua1798@gmail.com	Nữ	130000	
GV19	Iwasaki masahiro	Nhật Bản	miwasaki1960@gmail.com	Nữ	950	
GV20	Lê Hà Tuyết Ngân	Việt Nam	tuyetngan1998cnn@gmail.com	Nữ	110000	
GV21	Atsushi Harada	Nhật Bản	a.harada.4.30@gmail.com	Nữ	950	
GV22	Yasuhiro Kurosawa	Việt Nam	xkurosawa.yasuhiro@gmail.com	Nam	950	
GV23	Nguyễn Hải Quỳnh	Nhật Bản	nguyenhaiquynh97@gmail.com	Nữ	100000	
GV24	Nguyễn Thu Vân	Việt Nam	thuvannguyen2510@gmail.com	Nữ	100000	
GV25	Janyu Hitomi	Nhật Bản	hitohito7bababata@yahoo.co.jp	Nữ	950	
GV26	Yamashita Chisako	Nhật Bản	chisako@ac.auone-net.jp	Nữ	950	

☰	HocVien	HocVien_Salw	KhoaHoc	GiaoVien	NVSale	DangKy	ThanhToan	Hoc
---	---------	--------------	---------	----------	--------	--------	-----------	-----

- Thông tin dữ liệu khóa học:

	A	B	C	D	E	F	G
1	ID_KH	Ten	Sobuoi	LoaiKhoa	Hocphi	GiaKM	
2	SKAI11030JP	Kaiwa 1	30	1-1	10750000	8600000	
3	SKAI21030JP	Kaiwa 2	30	1-1	11875000	9500000	
4	SKAI31030JP	Kaiwa 3	30	1-1	12500000	10000000	
5	SKAI41030JP	Kaiwa 4	30	1-1	13125000	10500000	
6	SKAI51030JP	Kaiwa 5	30	1-1	13125000	10500000	
7	SKAI61030JP	Kaiwa 6	30	1-1	13125000	10500000	
8	CKAI121030JP	combo_kaiwa12	60	1-1	22625000	15500000	
9	CKAI231030JP	combo_kaiwa23	60	1-1	24375000	16500000	
0	CKAI341030JP	combo_kaiwa34	60	1-1	25625000	17400000	
1	CKAI451030JP	combo_kaiwa45	60	1-1	25925000	17800000	
2	CKAI561030JP	combo_kaiwa56	60	1-1	26250000	18200000	
3	CKAI131030JP	combo_kaiwa123	90	1-1	35125000	23500000	
4	CKAI141030JP	combo_kaiwa1234	120	1-1	48250000	32000000	
5	CKAI151030JP	combo_kaiwa12345	150	1-1	61375000	40000000	
6	CKAI241030JP	combo_kaiwa234	90	1-1	37500000	25000000	
7	CKAI251030JP	combo_kaiwa2345	120	1-1	50625000	33500000	
8	CKAI351030JP	combo_kaiwa345	90	1-1	38750000	26000000	
9	CKAI461030JP	combo_kaiwa456	90	1-1	42275000	27000000	
0	SKAIP11030VN	Pre-kaiwa1	20	1-1	6500000	4500000	
1	SKAIP21030VN	Pre-kaiwa2	15	1-1	6000000	4000000	
2	SKAI11030VN	Kaiwa 1	30	1-1	9000000	6500000	
3	SKAI21030VN	Kaiwa 2	30	1-1	10000000	7000000	
4	SKAI31030VN	Kaiwa 3	30	1-1	11000000	7500000	
5	SKAI41030VN	Kaiwa 4	30	1-1	13125000	10500000	
6	SKAI51030VN	Kaiwa 5	30	1-1	13125000	10500000	
7	CKAIP111030VN	combo_pre1+kaiwa1	50	1-1	13125000	10500000	
8	CKAIP211030VN	combo_pre2+kaiwa1	45	1-1	13125000	10500000	

☰	HocVien	HocVien_Salw	KhoaHoc	GiaoVien	NVSale	DangKy	ThanhToan
---	---------	--------------	---------	----------	--------	--------	-----------

- Thông tin dữ liệu học viên:

A	B	C	D	E	F
72019021	Trần thị mỹ duyên	Nữ	384593498	myduyen.tran.1806@gmail.com	Kaiwa 1
72019022	Trần Thị Hợi	Nữ	973615124	tranthihoi.08i1@gmail.com	Kaiwa 1
72019023	Nguyễn Thị Thúy	Nữ	981173000	naycoem@gmail.com	Kaiwa 2
72019024	Nguyen Thi Anh	Nữ	366125904	vanhtep1804@gmail.com	Kaiwa 1
72019025	Đặng Ngoan Cường	Nam	969808233	ngoancuong83@gmail.com	Kaiwa 1
72019026	Bùi Xuân Hiếu	Nam	965054001	xuanhieubuihd@gmail.com	Kaiwa 1
72019027	Kim Thị Yến	Nữ	966527719	bientapvienkimyen@gmail.com	Kaiwa 1
72019028	Duong anh dung	Nam	336957418	quang1958vinh@gmail.com	Kaiwa 1
72019029	Đỗ Văn Tùng	Nam	968162402	dovantung240994@gmail.com	Kaiwa 2
72019030	Phạm Thị Kiều Oanh	Nữ	819073190187	phamoanh.hy2002@gmail.com	Kaiwa 1
72019031	Nguyễn Thị Diễm An	Nữ	902732657	anan.ntd@gmail.com	Kaiwa 2
72019032	Trần Quốc Nghĩ	Nam	368300555	quocnghi.scv@gmail.com	Kaiwa 2
72019033	Cao Thị Thu Hương	Nữ	915139924	thuhuongcao86@gmail.com	Kaiwa 2
72019034	Cao Ngọc Nam	Nam	901666992	ngocnam7920@gmail.com	Kaiwa 2
72019035	Lê Trung Thế	Nam	374182328	letrungthe789@gmail.com	Kaiwa 1
72019036	Nguyễn Huy Dũng	Nam	915805996	huydung@vietinak.com.vn	Kaiwa 3
72019037	Đoàn Thị Vân	Nữ	916999557	vanpdu@gmail.com	Kaiwa 3
72019038	Hoàng Xuân Linh	Nam	964206964	hoangxuanlinhbkhnk58@gmail.com	Kaiwa 3
72019039	Vũ Đức Trung	Nam	7040863129	trungky86@gmail.com	Kaiwa 3
72019040	Hoàng Thị Phương Hà	Nữ	9056717111	hoanghache@gmail.com	Kaiwa 4
72019042	Phan Thị Anh Trâm	Nữ	972751807	phanthianhtram@gmail.com	Kaiwa 2
72019043	Lê Thị Êm	Nữ	966110618	saobien86.hd@gmail.com	Kaiwa 1
72019044	Nguyễn Thị Diễm	Nữ	1627386428	diemnguyen1095@gmail.com	Kaiwa 3
72019045	Trần Văn Phuộc	Nam	357423566	phuoc11tc313@gmail.com	Kaiwa 3
72019046	Bùi Mỹ Hạnh	Nữ	8081219223	hongmy14041998@gmail.com	Kaiwa 2
72019047	Nguyễn Hoàng Diệu	Nam	967806306	dieunguyenhoang359200@gmail.com	Kaiwa 2
72019048	Bùi Thị Lương	Nữ	988287785	buithiluong241988@gmail.com	Kaiwa 2
72019049	Chử Thị Thu Hương	Nữ	902387161	thuhuong9985@gmail.com	Kaiwa 1

HocVien	HocVien_Sale	KhoaHoc	GiaoVien	NVSale	DangKy	ThanhToan	Hoc
---------	--------------	---------	----------	--------	--------	-----------	-----

- Thông tin dữ liệu học viên trong phòng sale

A	B	C	D	E	F	G	H
ID_HV	Họ đệm	Tên	Gioitinh	Sđt	Email	TrinhDo	
72019001	Mi Xa	Tran	Female	917520969	mixa.brvt@gmail.com	Kaiwa 1	
72019002	Hoang Thi	Phuong	Female	842323225	phuongnguyen@gr...	Kaiwa 3	
72019003	Trần Việt	Hạnh	Female	914297679	viethanh.dn@gmail...	Kaiwa 1	
72019004	Nguyễn Thái	Hòa	Female	906220311	thaihoaftu@gmail...	Kaiwa 3	
72019005	Duy	Thành	Male	977701689	Duythanhdhc...	Kaiwa 5	
72019006	Duy	Bùi	Male	911263017	Builetuduy@gmail...	Kaiwa 2	
72019007	Trần Vĩnh	Hội	Male	386330057	vinhhoi2005@gmail...	Kaiwa 4	
72019008	Phạm Khánh	Ngoc	Female	914869998	khanhngoc0208@gmail...	Kaiwa 5	
72019009	Trần minh	trung	Male	869690927	tmtrung2010199@gmail...	Kaiwa 1	
72019010	Đức	Mạnh	Male	356442269	ng.manh.8888@gmail...	Kaiwa 3	
72019011	Vũ Thị Hằng	Nga	Female	978436072	sibobaka@gmail...	Kaiwa 3	
72019012	Lê Ngọc	Trinh	Female	764107695	lengoctrinh0312@gmail...	Kaiwa 2	
72019013	Vũ Văn	Anh	Male	368614246	vuvanken111894@gmail...	Kaiwa 3	
72019014	Võ Thị	Trà	Female	989539440	tra.vo2706@gmail...	Kaiwa 1	
72019015	Vũ Đinh	Phong	Male	8041811792	haphong210520@gmail...	Kaiwa 2	
72019016	Phí Văn	Hoàng	Male	903465700	apexoang@gmail...	Kaiwa 1	
72019017	Nguyễn Thị Tô	Lịch	Female	985306693	lequelinhchi168@gmail...	Kaiwa 1	
72019018	Nguyễn thị	thuỷ	Female	1699932708	phamquynhanh0@gmail...	Kaiwa 1	
72019019	Nguyễn Huỳnh Phương	Thảo	Female	903049822	nhpt017@gmail...	Kaiwa 1	
72019020	Phan Văn	Tuyễn	Male	817026884584	tuyen.japan2014@gmail...	Kaiwa 1	
72019021	Trần thị mỹ	duyên	Female	384593498	myduyen.tran.18@gmail...	Kaiwa 1	
72019022	Trần Thị	Hội	Female	973615124	tranthihoi.08i1@gmail...	Kaiwa 1	
72019023	Nguyễn Thị	Thúy	Female	981173000	naycoem@gmail...	Kaiwa 2	
72019024	Nguyen Thi	Anh	Female	366125904	vanhtep1804@gmail...	Kaiwa 1	
72019025	Đặng Ngoan	Cường	Male	969808233	ngoancuong83@gmail...	Kaiwa 1	
72019026	Bùi Xuân	Hiếu	Male	965054001	xuanhieubuihd@gmail...	Kaiwa 1	

☰	HocVien	HocVien_Sales	KhoaHoc	GiaoVien	NVSale	DangKy	ThanhToan
---	---------	---------------	---------	----------	--------	--------	-----------

Phụ lục 2: Hướng dẫn các thao tác cài đặt và kết nối với dịch vụ tích hợp máy chủ SQL (SQL Server Integration Services) khi sử dụng MySQL.

Chuẩn bị:

- MySQL Server & MySQL Workbench 8.0 CE
- ODBC data sources (32bit)

Lưu ý: SQL Server Integration Services chỉ hỗ trợ ODBC data sources bản 32bit

Cách kết nối:

Bước 1: Tạo database và user



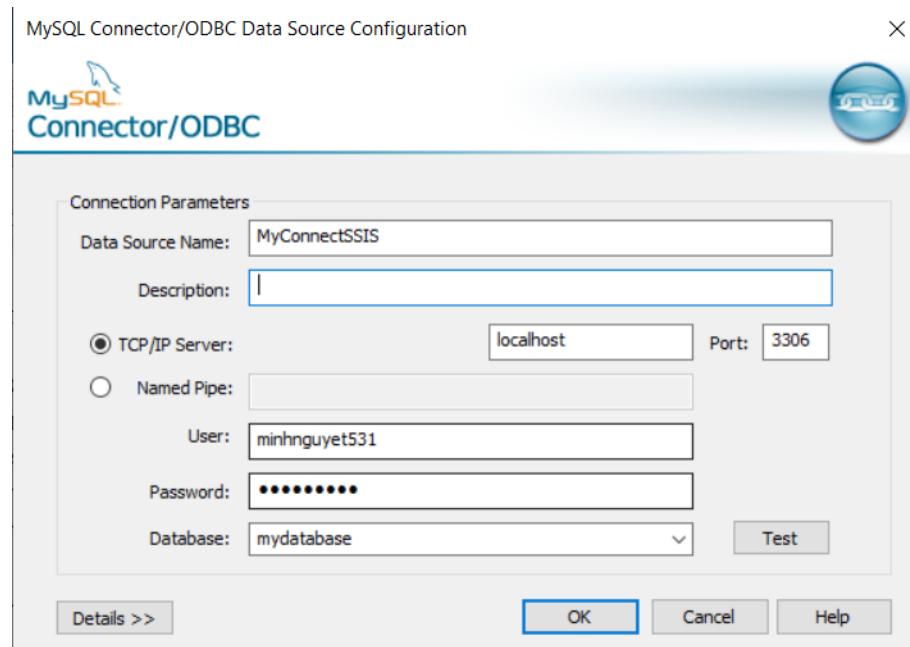
```

1 1. create database <tên database>;
2   vd: create database mydatabase;
3
4 2. create user '<Tên user>'@'localhost' identified by '<password>';
5   vd: create user 'minhnguyet531'@'localhost' identified by '123456789';
6
7 3. grant all on <tên database>.* to '<Tên user>'@'localhost';
8   vd: grant all on mydatabase.* to 'data'@'localhost';
9
10 4. flush privileges;
11 5. show grants for '<Tên user>'@'localhost';
12   vd: show grants for 'minhnguyet531'@'localhost';

```

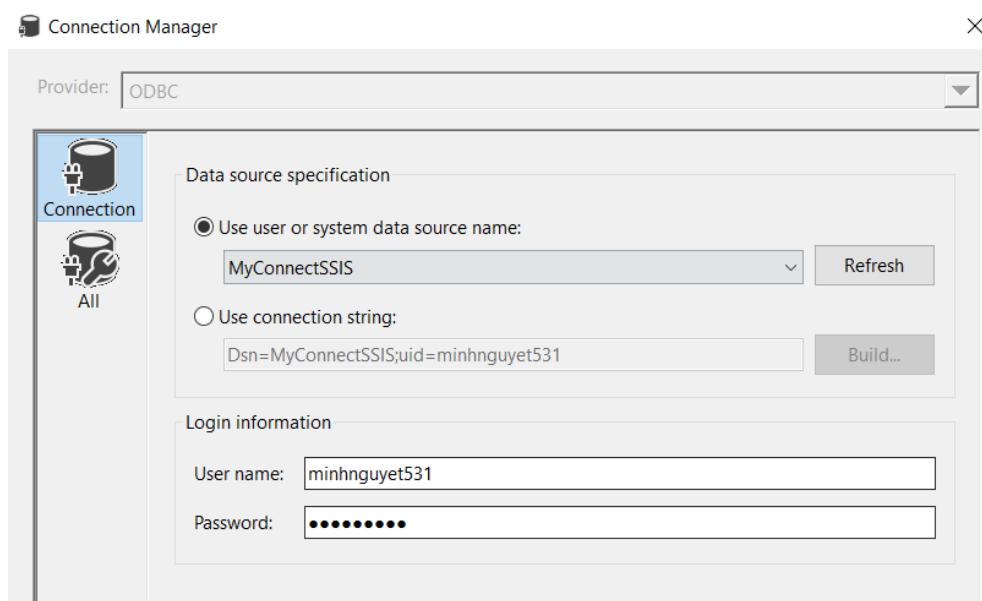
Bước 2: Tạo User Data sources

- Mở tool ODBC data sources (32bit) tại tab User DSN → Add → Chọn MySQL ODBC 8.0 Unicode Driver → Finish
- Điền thông tin tương ứng (Tại phần **Data Source Name**: Đặt tên ngẫu nhiên. **User/Pass**: Điền thông tin user có quyền được thao tác với database sau đó chọn trường database mk muốn thao tác.) → **OK**



Bước 3: Thêm nguồn kết nối với SSIS

Từ **Configure ODBC Connection Manager** nhấn **New...** Tại phần **Connection Manager** lựa chọn tên **User Data sources** vừa tạo được ở bước 1. Nhập **User name/ Password** tương tác được với database đã tạo. → **OK**



Cách sử dụng: Sau khi hoàn thành 2 bước trong phần trên chúng ta có thể kết nối với SSIS tương tự như các cách với SQL server. Lưu ý khi đó nguồn vào của data phải là **ODBC Source** và đích của nó là **ODBC Destination**.

Phụ lục 3: Khai phá Dữ liệu

Phần code Khai phá dữ liệu Dự đoán khả năng học viên tiếp tục mua tiếp khóa học tại trung tâm:

<https://github.com/minhnguyet531/DataMining/blob/main/HeHoTroCayQuyetDinh.ipynb>

File dataset mà nhóm sử dụng:

<https://github.com/minhnguyet531/DataMining/blob/main/datasetHeHoTro.xlsx>